*mathematics*

MDPI

*Article*

# Generalizations and Properties of Normalized Similarity Measures for Boolean Models

**Amelia Bădică [1], Costin Bădică [2,*], Doina Logofătu [3] and Ionuţ-Dragoş Neremzoiu [2]**

[1] Department of Statistics and Business Informatics, University of Craiova, 200585 Craiova, Romania; amelia.badica@edu.ucv.ro
[2] Department of Computers and Information Technology, University of Craiova, 200440 Craiova, Romania; neremzoiu.ionut.k4c@student.ucv.ro
[3] Faculty of Computer Science and Engineering, Frankfurt University of Applied Sciences, Nibelungenplatz 1, 60318 Frankfurt am Main, Germany; logofatu@fb2.fra-uas.de
* Correspondence: costin.badica@edu.ucv.ro

**Abstract:** In this paper, we provide a closer look at some of the most popular normalized similarity/distance measures for Boolean models. This work includes the generalization of three classes of measures described as generalized Kulczynski, generalized Jaccard, and generalized Consonni and Todeschini measures, theoretical ordering of the similarity measures inside each class, as well as between classes, and positive and negative results regarding the metric properties of measures related to satisfying or not satisfying the triangle inequality axiom.

**Keywords:** distance measure; similarity measure; metric; Boolean model; finite set

**MSC:** 03E75; 68R01; 68T01

## 1. Introduction

Defining a quantitative measure that evaluates the similarity of two objects occurs in various scientific areas, including computing, natural sciences, medicine, forensics, socio-economic sciences, engineering, and arts. For example, biologists use similarity measures for sequence alignment, protein structure comparison, and studying the similarity in species distribution. Computer scientists are using similarity measures for comparing data points in clustering, classification, and anomaly detection, for determining relevant documents in information retrieval, for assessing semantic compatibility of natural language texts, as well as for object and speech recognition in computer vision and multi-modal user interfaces. Medical doctors are interested in assessing medical image similarity in radiology or evaluating disease spread patterns in epidemiology. Sociologists and psychologists are conducting survey response analysis and evaluating behavioral similarity, while forensic engineers are analyzing fingerprints and handwriting. Last but not least, economists are assessing the similarity of consumer behaviors and stock patterns, artists and musicians are interested in melody similarity and art styles, while Earth scientists are analyzing terrain similarity, map matching, and weather patterns.

In this paper, we provide a theoretical investigation of some of the most popular similarity/distance measures for Boolean models. For a fair discussion and comparison we restrict our attention to normalized similarities (i.e., bounded by $[0, 1]$) of type 1 according to [1], which only depend on characteristics present in the compared objects. In particular, we are interested in developing clear definitions and generalizations of the measures, as well as analyzing their theoretical ordering and metric properties in detail.

check for updates

Academic Editor: Óscar Valero Sierra

Received: 6 January 2025
Revised: 21 January 2025
Accepted: 22 January 2025
Published: 24 January 2025

**Citation:** Bădică, A.; Bădică, C.; Logofătu, D.; Neremzoiu, I.-D. Generalizations and Properties of Normalized Similarity Measures for Boolean Models. *Mathematics* **2025**, *13*, 384. https://doi.org/10.3390/math13030384

We provide new theoretical results covering the following three aspects:

i. Generalization of three classes of measures described as generalized Kulczynski, generalized Jaccard, and generalized Consonni and Todeschini measures (Definition 7, Proposition 6, Equation (36)).
ii. Theoretical ordering of the similarity measures inside each class, as well as between classes (Proposition 4, Proposition 6, Proposition 7).
iii. Positive and negative results regarding the metric properties of the measures related to satisfying or not satisfying the triangle inequality axiom (Proposition 5, Proposition 8, Proposition 9, Proposition 10).

To the best of our knowledge, the proposed generalizations of similarity measures, as well as the results regarding the theoretical ordering, are new. Moreover, the positive and negative results regarding satisfying or not satisfying the triangle inequality axiom can be seen as generalizations of particular results concerning specific similarity measures, like the metricity of Jaccard similarity and non-metricity of cosine similarity.

The paper starts with a brief overview of related works in Section 2. Section 3 covers background definitions of similarity and distance measures, metrics, and set measures, as well as notations used in the rest of the paper. Section 4 introduces the most popular normalized similarity measures found in the literature. Sections 5–7 cover the most important results regarding the generalization of Kulczynski, Jaccard, and Consonni and Todeschini similarity measures and their properties.

## 2. Related Works

The Jaccard similarity was proposed more than one century ago in [2] to study patterns of flora distribution. The first mathematical proof that Jaccard distance satisfies triangle inequality was provided many decades later in [3]. One year later, a shorter and very elegant proof was provided by [4]. Two recent generalizations of Jaccard similarity using modular and submodular set functions, including metricity proofs, were provided in [5]. It is interesting to note that these results apply to general distributive lattices rather than just sets. Note, however, that they are different from our proposal of a parameterized version of Jaccard similarity as a symmetric Tverski index.

A very recent axiomatic characterization and generalization of the Jaccard similarity metric is provided in [6], featuring the increase in marginal sensitivity to the gradual removal of common elements (i.e., updating axiom A3 from [6]). It is interesting to note that the cited study shows by means of a simple example that their proposed generalization of Jaccard (different from ours), Sørensen–Dice, cos (i.e., Driver and Kroeber or Ochiai), and overlap similarities do not satisfy triangle inequality (axiom A7 in [6]). However, our results are more general, stating that the generalized Kulczynski distance satisfies triangle inequality only for the particular Braun–Blanquet case and fails to satisfy it for all other mean functions.

An early overview of binary, i.e., presence–absence and similarity coefficients, is provided in [7]. While the authors observed that most of the binary coefficients considered for discussion were developed intuitively and tested empirically, their purpose was to highlight their conceptual relationships and to standardize their symbolic expressions. Nevertheless, other theoretical results are missing.

An evaluation of 43 coefficients of association and similarity based on binary data is proposed in [8]. That analysis involves theoretical considerations of admissibility and other additional conditions, as well as an empirical evaluation of the significance of their association on a real data basis using the chi-square test. An interesting conclusion of that work was that "a set of measures that generally work well" should comprise Sørensen–Dice, Kulczynski, Driver and Kroeber (Ochiai), and Braun–Blanquet similarities. According

to our analysis, these four measures are, in fact, special cases of generalized Kulczynski similarity, thus theoretically explaining some of the experimental conclusions of [8].

An experimental comparison of the effect of base rates on the grouping of 71 binary similarity coefficients was recently provided in [9]. However, while this comparison involves a much larger ecosystem of similarity measures for binary models than our work, the results obtained are purely experimental and address a very specific and practical problem rather than investigating the more general theoretical properties of the analyzed measures, like ordering and metricity.

A recent survey of similarity measures for both binary and numerical data is presented in [1]. While the results of this survey address both binary and numerical data, the analysis provided covers mainly semantic aspects: order-based comparison, value-based comparison, degree of severity, and power of discrimination, while other mathematical properties that are more relevant for stronger theoretical results, like metricity properties, are not investigated.

Yet another path of approaching the definition of distances between finite sets is to regard them as capturing discrete probability distributions. Probably the earliest work in this direction is provided by Rajski's distance [10], which is defined based on Shannon entropy and joint entropy [11]. Rajski proved that his distance defines a metric space of discrete probability distributions, thus satisfying triangle inequality. An abstract definition of entropy and its metricity properties in the context of submodular and supermodular functions on lattices is proposed in [12]. These results could provide an interesting basis for investigating extensions of our work, covering the probabilistic aspect in the framework of general lattices.

Moreover, rather more recently, a new measure of similarity of Boolean vectors (i.e., finite sets), defined as the collision probability of optimal locality-sensitive hashing, was proposed in [13]. Their proposed probability Jaccard similarity is shown to be a natural generalization of the Jaccard similarity to probability distributions, as compared to the weighted Jaccard index. Interesting connections between locality-sensitive hashing and supermodularity of the similarity measures are proposed in [14]. This might suggest possible connections between their results and our results on generalized Consonni and Todeschini measures relying on the supermodularity property.

## 3. Background

A similarity measure or index, or simply similarity, at its core, quantifies the similarity between two (or more) objects. In what follows, we address only the binary case.

The most general axiomatic definition of similarity is provided below [15].

**Definition 1.** *(Similarity Measure) Let $X$ be a set of objects. A function $s : X \times X \to \mathbb{R}$ is called a similarity measure on $X$ if and only if it satisfies the following axioms:*

1. *It is non-negative, i.e., $s(x, y) \geq 0$ holds for all $x, y \in X$.*
2. *It is symmetric, i.e., $s(x, y) = s(y, x)$ holds for all $x, y \in X$.*
3. *It satisfies the identity of indiscernibles, i.e., $s(x, y) \leq s(x, x)$ holds for all $x, y \in X$.*

A distance measure or simply distance, also known as dissimilarity, at its core, quantifies the dissimilarity between two (or more) objects. Observe that distance is exactly the opposite of similarity, i.e., whenever the similarity is higher, the distance is lower and vice versa.

**Definition 2.** *(Distance measure) A function $d : X \times X \to \mathbb{R}$ is called a distance (or dissimilarity) measure on $X$ if it satisfies the following axioms:*

1. *It is non-negative, i.e.,* $d(x,y) \geq 0$ *holds for all* $x, y \in X$.
2. *It is symmetric, i.e.,* $d(x,y) = d(y,x)$ *holds for all* $x, y \in X$.
3. *It is reflexive, i.e.,* $d(x,x) = 0$ *holds for all* $x, y \in X$.

Very often, distance measures are required to reflect the real-world geometric intuition that the direct distance between two points is always shorter than or equal to the sum of distances through an intermediate point. This constraint is known as triangle inequality, and it has the advantage, on one hand, of interpretability and consistency with human reasoning, and on the other hand, of providing some logical properties that are necessary for the mathematical foundations of many algorithms and theories.

**Definition 3.** *((Semi)Metric) A distance measure* $d : X \times X \to \mathbb{R}$ *is semimetric on* $X$ *if it satisfies the triangle inequality axiom:*

1. $d(x,y) \leq d(x,z) + d(z,y)$ *for all* $x, y, z \in X$.
2. *Additionally,* $d(x,y) = 0$ *if and only if* $x = y$ *for all* $x, y \in X$.

*Then,* $d$ *is called a metric.*

Given a general similarity measure, one can easily build a corresponding distance measure as follows.

**Proposition 1.** *(Distance corresponding to a similarity) If* $s : X \times X \to \mathbb{R}$ *is a similarity then* $d : X \times X \to \mathbb{R}$ *defined as follows:*

$$d(x,y) = \min\{s(x,x), s(y,y)\} - s(x,y), \tag{1}$$

*where it is a distance.*

**Proof.** We show that $d(\cdot,\cdot)$ defined by Equation (1) satisfies the distance axioms of Definition 2.

From item 3 of Definition 1 it follows that $s(x,x) \geq s(x,y)$ and $s(y,y) \geq s(y,x)$. Using the symmetry of $s(\cdot,\cdot)$ (item 2, Definition 1) we get the positiveness of $d(\cdot,\cdot)$.

Also, from the symmetry of $s(\cdot,\cdot)$, we obtain $d(x,y) = \min\{s(x,x), s(y,y)\} - s(x,y) = \min\{s(y,y), s(x,x)\} - s(y,x) = d(y,x)$, i.e., proving $d(\cdot,\cdot)$ is symmetric.

Finally, $d(x,x) = \min\{s(x,x), s(x,x)\} - s(x,x) = s(x,x) - s(x,x) = 0$, concluding the proof. □

The similarity function $s$ is very often normalized, i.e., it is bound to the interval $[0,1]$, and the similarity of each element $x \in X$ to itself is 1; that is, $s(x,x) = 1$ for all $x \in X$. In this way, axioms 1 and 3 of Definition 1 are automatically satisfied.

**Definition 4.** *(Normalized similarity measure) Let* $X$ *be a set of objects. Function* $s : X \times X \to [0,1]$ *is called a normalized similarity measure on* $X$ *if it satisfies the following axioms:*

1. *It is symmetric, i.e.,* $s(x,y) = s(y,x)$ *holds for all* $x, y \in X$.
2. $s(x,x) = 1$ *for all* $x \in X$.

It is obvious that a normalized similarity measure is a similarity measure. Moreover, according to Proposition 1, if $s$ is a normalized similarity, then $d(x,y) = 1 - s(x,y)$ is a normalized distance, i.e., $d(x,y) \in [0,1]$ for all $x, y \in X$.

In what follows, we focus on similarities and distances for objects described as sets of features, characteristics, or attributes. Moreover, we stick to the simplest Boolean model, i.e., each feature can be present or absent in a given object.

Let us consider the finite universal set $\mathcal{U}$. This set can be interpreted as the universal set of characteristics or features that can be possessed by or observed at an object. According to the Boolean model, a feature $u \in \mathcal{U}$ can be present or not in a given object or unit of observation. So, we can represent an object by its set of features, denoted as $\mathcal{X} \subseteq \mathcal{U}$.

We carefully analyze some of the most important and popular normalized similarities proposed in the literature for objects described using the Boolean model. Adhering to the normalized model has the advantage that various similarities can be compared, allowing us to theoretically establish a domain-independent ordering of some of the most well-known normalized similarities.

It is not difficult to see that, according to the notation from Definition 4, the set of objects $X$ in the Boolean model is actually the power set $2^{\mathcal{U}}$, while the objects themselves correspond to subsets of $\mathcal{U}$.

For each set $\mathcal{X} \subseteq \mathcal{U}$, its complement set is defined as $\overline{\mathcal{X}} = \mathcal{U} \setminus \mathcal{X}$. Moreover, we denote $|\mathcal{X}|$ as the cardinal (number of elements) of $\mathcal{X}$.

Similarity involves two objects; let us denote them by subsets $\mathcal{X}$ and $\mathcal{Y}$ of $\mathcal{U}$.

First, observe the following:

i. $\mathcal{X} \setminus \mathcal{Y}, \mathcal{X} \cap \mathcal{Y}$ is a partition of $\mathcal{X}$, and symmetrically, $\mathcal{Y} \setminus \mathcal{X}, \mathcal{X} \cap \mathcal{Y}$ is a partition of $\mathcal{Y}$.
ii. $\mathcal{X} \setminus \mathcal{Y}, \mathcal{Y} \setminus \mathcal{X}, \mathcal{X} \cap \mathcal{Y}$ is a partition of $\mathcal{X} \cup \mathcal{Y}$.
iii. $\mathcal{X} \setminus \mathcal{Y}, \mathcal{Y} \setminus \mathcal{X}$ is a partition of the symmetric difference $\mathcal{X} \Delta \mathcal{Y}$.
iv. $\mathcal{X} \setminus \mathcal{Y}, \mathcal{Y} \setminus \mathcal{X}, \mathcal{X} \cap \mathcal{Y}, \overline{\mathcal{X} \cup \mathcal{Y}}$ is a partition of $\mathcal{U}$.

Standard definitions of similarities are given in terms of the cardinal. However, in what follows, we generalize these definitions for arbitrary measures on finite sets.

**Definition 5.** *(Set measure) A measure on finite sets is a function $\mu : 2^{\mathcal{U}} \to \mathbb{R}$ that satisfies the following axioms:*

*i. It is positive, i.e., $\mu(\mathcal{X}) \geq 0$ for each set $\mathcal{X} \subseteq \mathcal{U}$.*
*ii. It is additive for disjoint sets, i.e., $\mu(\mathcal{X} \cup \mathcal{Y}) = \mu(\mathcal{X}) + \mu(\mathcal{Y})$ for disjoint sets $\mathcal{X}, \mathcal{Y} \subseteq \mathcal{U}$.*

A set measure $\mu$ has some properties that are very useful in the theoretical analysis of similarity measures based on $\mu$.

**Proposition 2.** *Let $\mu$ be a measure on finite set $\mathcal{U}$. Then, the following hold:*

*i. The empty set has a null measure, i.e., $\mu(\emptyset) = 0$.*
*ii. It is monotone, i.e., $\mu(\mathcal{X}) \leq \mu(\mathcal{Y})$ for $\mathcal{X} \subseteq \mathcal{Y}$.*
*iii. $\mu$ is modular, i.e., $\mu(\mathcal{X} \cup \mathcal{Y}) + \mu(\mathcal{X} \cap \mathcal{Y}) = \mu(\mathcal{X}) + \mu(\mathcal{Y})$.*
*iv. $\mu$ is non-negative and bounded, i.e., $0 \leq \mu(\mathcal{X}) \leq \mu(\mathcal{U})$ for all $\mathcal{X} \subseteq \mathcal{U}$.*

**Proof.** Observe that from additivity, it follows that $\mu(\emptyset) + \mu(\emptyset) = \mu(\emptyset)$, so $\mu(\emptyset) = 0$.

Considering $\mathcal{X} \subseteq \mathcal{Y}$, it follows that $(\mathcal{X}, \mathcal{Y} \setminus \mathcal{X})$ is a partition of $\mathcal{Y}$. Then, from additivity, $\mu(\mathcal{Y}) = \mu(\mathcal{X}) + \mu(\mathcal{Y} \setminus \mathcal{X})$. Using positivity $\mu(\mathcal{Y} \setminus \mathcal{X}) \geq 0$, it follows that $\mu(\mathcal{X}) \leq \mu(\mathcal{Y})$, resulting the monotony of $\mu(\cdot)$.

Modularity follows from additivity and partitioning $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \cup \mathcal{Y}$ in terms of $\mathcal{X} \setminus \mathcal{Y}$, $\mathcal{Y} \setminus \mathcal{X}$, and $\mathcal{X} \cap \mathcal{Y}$.

From monotony, as $\mathcal{X} \subseteq \mathcal{U}$, it follows that $\mu(\mathcal{X}) \leq \mu(\mathcal{U})$ for all $\mathcal{X} \subseteq \mathcal{U}$. □

Moreover, it is not difficult to observe that measures of finite sets have a general form.

**Proposition 3.** *$\mu$ is a measure on finite set $\mathcal{U}$ if and only if there exist constants $m_i \geq 0$ for each $i \in \mathcal{U}$, such that:*

$$\mu(\mathcal{X}) = \sum_{i \in \mathcal{X}} m_i \quad \text{for all } \mathcal{X} \subseteq \mathcal{U}. \tag{2}$$

**Proof.** It is very easy to observe that if $\mu$ satisfies Equation (2), then $\mu$ satisfies axioms of Definition 5. Conversely, if $\mu$ satisfies these axioms, then letting $m_i = \mu(\{i\})$ for all $i \in \mathcal{U}$, we obtain from the positivity axiom that $m_i \geq 0$ and from the additivity axiom that $\mu$ satisfies Equation (2). □

## 4. Similarity Measures

The definition of the most important and popular similarity measures is often given in terms of the following notations:

$$\begin{aligned} a &= \mu(\mathcal{X} \cap \mathcal{Y}) \text{ attributes common to } \mathcal{X} \text{ and } \mathcal{Y}, \\ b &= \mu(\mathcal{X} \setminus \mathcal{Y}) \text{ attributes common to } \mathcal{X} \text{ only (and absent from } \mathcal{Y}\text{)}, \\ c &= \mu(\mathcal{Y} \setminus \mathcal{X}) \text{ attributes common to } \mathcal{Y} \text{ only (and absent from } \mathcal{X}\text{)}, \\ d &= \mu(\overline{\mathcal{X} \cup \mathcal{Y}}) \text{ attributes absent in both } \mathcal{X} \text{ and } \mathcal{Y}. \end{aligned} \tag{3}$$

Note that, as $\mathcal{X} \cap \mathcal{Y}, \mathcal{X} \setminus \mathcal{Y}, \mathcal{Y} \setminus \mathcal{X}, \overline{\mathcal{X} \cup \mathcal{Y}}$ is a partition of $\mathcal{U}$ and $\mu(\cdot)$ is a measure, the following hold:

$$\begin{aligned} \mu(\mathcal{U}) &= a + b + c + d, \\ \mu(\mathcal{X}) &= a + b, \\ \mu(\mathcal{Y}) &= a + c, \\ \mu(\mathcal{X} \Delta \mathcal{Y}) &= b + c, \\ \mu(\mathcal{X} \cup \mathcal{Y}) &= a + b + c. \end{aligned} \tag{4}$$

As already pointed out in Section 2, there are so many similarities proposed in the literature with applications in very diverse areas. For a fair discussion and comparison, we are going to restrict our attention to normalized similarities (i.e., those satisfying axioms of Definition 4).

According to [1], type 1 similarity measures are those that only depend on characteristics present either in $\mathcal{X}$ or $\mathcal{Y}$ (possibly in both, i.e., $a$, $b$, and $c$) but are independent of the attributes absent of both objects (i.e., they do not depend on $d$). Type 2 similarity measures are those that take into account all four quantities derived from the objects, i.e., their intersection, set differences, and the intersection of their complementary sets (i.e., $a$, $b$, $c$, and $d$). The major difference with type 1 similarity measures is that for type 2 measures, the size of the universe influences similarity. Consequently, depending on the measures, two objects can be more similar in a smaller universe than in a larger one, as "measured" by $\mu$.

In what follows, we will focus on type 1 similarities, i.e., those that depend solely on $\mathcal{X}$ and $\mathcal{Y}$, and not on the features found in other objects of the data set, i.e., not on features in $\overline{\mathcal{X} \cup \mathcal{Y}}$. In other words, we focus on those similarities depending only on $a, b, c$ and not on $d$. Note that our assumption corresponds to the adoption of the "matching" axiom of similarity measures according to [16].

We are going to provide the defining equations of similarities in terms of $a, b, c$ and also in terms of $\mathcal{X}, \mathcal{Y}$. Moreover, all the definitions from the literature consider similarity measures based on set cardinality, $\mu(\cdot) \equiv |\cdot|$. In what follows, we provide more general definitions for arbitrary measures $\mu(\cdot)$.

### *4.1. Jaccard Similarity*

The Jaccard similarity index [2] is defined as follows:

$$J(\mathcal{X}, \mathcal{Y}) = \frac{\mu(\mathcal{X} \cap \mathcal{Y})}{\mu(\mathcal{X} \cup \mathcal{Y})} = \frac{a}{a + b + c}. \tag{5}$$

### 4.2. Sørensen–Dice Similarity

The Sørensen–Dice similarity index [17,18] is defined as follows:

$$SD(\mathcal{X},\mathcal{Y}) = \frac{2\mu(\mathcal{X}\cap\mathcal{Y})}{\mu(\mathcal{X})+\mu(\mathcal{Y})} = \frac{2a}{2a+b+c}. \tag{6}$$

### 4.3. Ochiai/Cos Similarity

The Driver and Kroeber [19] or Ochiai index [20] is defined as follows:

$$DKO(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{\sqrt{\mu(\mathcal{X})\mu(\mathcal{Y})}} = \frac{a}{\sqrt{(a+b)(a+c)}}. \tag{7}$$

Note that if $\mu(\cdot) \equiv |\cdot|$, then the Ochiai index is the same as the cosinus of the Boolean vectors representing sets $\mathcal{X}$ and $\mathcal{Y}$. That is why the Ochiai index is sometimes called cos similarity. Equation (7) can be also seen as an immediate generalization of the cos similarity to arbitrary measures.

### 4.4. Sorgenfrei Similarity

The Sorgenfrei similarity index [21] is defined as follows:

$$SO(\mathcal{X},\mathcal{Y}) = \frac{(\mu(\mathcal{X}\cap\mathcal{Y}))^2}{\mu(\mathcal{X})\mu(\mathcal{Y})} = \frac{a^2}{(a+b)(a+c)}. \tag{8}$$

Observe the following:

$$SO(\mathcal{X},\mathcal{Y}) = (DKO(\mathcal{X},\mathcal{Y}))^2. \tag{9}$$

### 4.5. Sokal and Sneath Similarity

The Sokal and Sneath similarity index [22] is defined as follows:

$$SS(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{2\mu(\mathcal{X}\cup\mathcal{Y})-\mu(\mathcal{X}\cap\mathcal{Y})} = \frac{a}{a+2(b+c)}. \tag{10}$$

### 4.6. Kulczynski Similarity

The Kulczynski similarity index [23] is defined as follows:

$$KU(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{2}\left(\frac{1}{\mu(\mathcal{X})}+\frac{1}{\mu(\mathcal{Y})}\right) = \frac{a}{2}\left(\frac{1}{a+b}+\frac{1}{a+c}\right). \tag{11}$$

### 4.7. Overlap or Szymkiewicz—Simpson Similarity

The overlap or Szymkiewicz—Simpson similarity index appears as formula #27 in Table 2 from [24], as well as in [25], and it is defined as follows:

$$OV(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{\min(\mu(\mathcal{X}),\mu(\mathcal{Y}))} = \frac{a}{\min(a+b,a+c)}. \tag{12}$$

### 4.8. Braun–Blanquet Similarity

The Braun–Blanquet similarity index [26] is defined as follows:

$$BB(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{\max(\mu(\mathcal{X}),\mu(\mathcal{Y}))} = \frac{a}{\max(a+b,a+c)}. \tag{13}$$

### *4.9. Consonni and Todeschini Similarity*

The Consonni and Todeschini similarity index [27] is defined as follows:

$$CT(\mathcal{X},\mathcal{Y}) = \frac{\log(1+\mu(\mathcal{X}\cap\mathcal{Y}))}{\log(1+\mu(\mathcal{X}\cup\mathcal{Y}))} = \frac{\log(1+a)}{\log(1+a+b+c)}. \tag{14}$$

**Example 1.** *Let us consider $n \in \mathbb{N}^*$ and the following three sets:*

$$\begin{aligned} \mathcal{X} &= \{3,6,\ldots,3n\}, \\ \mathcal{Y} &= \{1,2,3,\ldots,3n-2,3n-1,3n\} = \mathcal{X}\cup\mathcal{Z}. \end{aligned} \tag{15}$$

*Let us consider that all the elements $2,3,\ldots,3n$ have measure 1, while element 1 has measure 2. We trivially obtain $a = n$, $b = 0$, and $c = 2n+1$ using Equation (3) for sets $\mathcal{X}$ and $\mathcal{Y}$. Computing similarity measures with Equations (5)–(8), (10)–(14) for sets $\mathcal{X}$ and $\mathcal{Y}$, we obtain the results presented in Table 1.*

**Table 1.** Similarity values for sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1.

| **General values for $n \in \mathbb{N}^*$** | **Values for $n = 1$** | **Values for $n = 2$** |
|---|---|---|
| $J = BB = SO = \frac{n}{3n+1}$ | $J = BB = SO = 0.250$ | $J = BB = SO = 0.285$ |
| $SD = \frac{2n}{4n+1}$ | $SD = 0.400$ | $SD = 0.444$ |
| $DKO = \sqrt{\frac{n}{3n+1}}$ | $DKO = 0.500$ | $DKO = 0.534$ |
| $SS = \frac{n}{5n+2}$ | $SS = 0.142$ | $SS = 0.166$ |
| $KU = \frac{4n+1}{6n+2}$ | $KU = 0.625$ | $KU = 0.642$ |
| $OV = 1.000$ | $OV = 1.000$ | $OV = 1.000$ |
| $CT = \frac{\log(n+1)}{\log(3n+2)}$ | $CT = 0.439$ | $CT = 0.528$ |

## 5. Generalized Kulczynski Similarity

### *5.1. Definition and Basic Properties*

Sørensen–Dice (6), Ochiai (7), Kulczynski (11), overlap (12), and Braun–Blanquet (13) similarities can be generalized to a unique similarity measure parameterized by a suitable mean function.

**Definition 6.** *(Mean function) Function $f : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ is called the mean if it satisfies the following conditions:*

*i. Boundedness:* $\min(x,y) \le f(x,y) \le \max(x,y)$ *for all* $x,y \in \mathbb{R}^+$.
*ii. Symmetry:* $f(x,y) = f(y,x)$ *for all* $x,y \in \mathbb{R}^+$.
*iii. Homogeneity with multiplication:* $f(\alpha x,\alpha y) = \alpha f(x,y)$ *for all* $\alpha,x,y \in \mathbb{R}^+$.

Classical examples of mean functions are as follows [28]:

$$\begin{aligned} \text{Arithmetic mean: } AM(x,y) &= \frac{x+y}{2}, \\ \text{Geometric mean: } GM(a,y) &= \sqrt{xy}, \\ \text{Harmonic mean: } HM(x,y) &= \frac{2xy}{x+y}, \\ \text{Power mean: } PM_p(x,y) &= \left(\frac{x^p+y^p}{2}\right)^{\frac{1}{p}} \text{ for } p \in \mathbb{R}^*, \\ \text{Logarithmic mean: } LM(x,y) &= \frac{y-x}{\log(y)-\log(x)}. \end{aligned} \tag{16}$$

Trivial examples of mean functions are $\max(x, y)$ and $\min(x, y)$.

Note that from the first point of Definition 6, it follows that $x = \min(x, x) \leq f(x, x) \leq \max(x, x) = x$, so $f(x, x) = x$ for all $x \in \mathbb{R}^+$, i.e., $f$ is idempotent.

We can now define the generalized Kulczynski similarity as follows:

**Definition 7.** *Let $f$ be an arbitrary mean function. Generalized Kulczynski similarity is defined by the following:*

$$GKU_f(\mathcal{X}, \mathcal{Y}) = \frac{\mu(\mathcal{X} \cap \mathcal{Y})}{f(\mu(\mathcal{X}), \mu(\mathcal{Y}))} = \frac{a}{f(a+b, a+c)}. \tag{17}$$

Note that the computational overhead of $GKU_f$ compared to the specific similarities (6), (7), (11), (12), and (13) is given solely by the computation of the mean function $f$.

**Proposition 4.**

*i. Generalized Kulczynski similarity (17) is a generalization of Sørensen–Dice (6), Ochiai (7), Kulczynski (11), overlap (12), and Braun–Blanquet (13) similarities as follows:*

$$\begin{aligned} SD(\mathcal{X}, \mathcal{Y}) &= GKU_{AM}(\mathcal{X}, \mathcal{Y}), \\ KU(\mathcal{X}, \mathcal{Y}) &= GKU_{HM}(\mathcal{X}, \mathcal{Y}), \\ DKO(\mathcal{X}, \mathcal{Y}) &= GKU_{GM}(\mathcal{X}, \mathcal{Y}), \\ OV(\mathcal{X}, \mathcal{Y}) &= GKU_{\min}(\mathcal{X}, \mathcal{Y}), \\ BB(\mathcal{X}, \mathcal{Y}) &= GKU_{\max}(\mathcal{X}, \mathcal{Y}). \end{aligned} \tag{18}$$

*ii.*

$$SO(\mathcal{X}, \mathcal{Y}) \leq BB(\mathcal{X}, \mathcal{Y}) \leq SD(\mathcal{X}, \mathcal{Y}) \leq DKO(\mathcal{X}, \mathcal{Y}) \leq KU(\mathcal{X}, \mathcal{Y}) \leq OV(\mathcal{X}, \mathcal{Y}). \tag{19}$$

**Proof.** The first point follows by direct substitution from the defining equations of the similarities.

The second point follows almost entirely (except the first inequality) from Equation (18) and from the well-known ordering of the classical mean functions [28]:

$$\max(a, b) \geq AM(a, b) \geq GM(a, b) \geq HM(a, b) \geq \min(a, b). \tag{20}$$

We only have to prove the following:

$$\begin{aligned} &SO(\mathcal{X}, \mathcal{Y}) \leq BB(\mathcal{X}, \mathcal{Y}), \\ &\text{i.e., } a \max(a+b, a+c) \leq (a+b)(a+c). \end{aligned}$$

The last inequality follows by observing that $\max(a+b, a+c)$ is $a+b$ or $a+c$, $a \leq a+c$, and $a \leq a+b$. □

**Example 2.** *It is not difficult to observe that Inequality (19) is satisfied for sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1 by simple algebraic manipulation (first column of Table 1) and simple observation (second and third columns of Table 1).*

### *5.2. Metric Properties*

In almost all the cases, the distance measure derived from generalized Kulczynski similarity does not satisfy the triangle inequality with one notable exception: the Braun–Blanquet case. These results are established by the following proposition.

**Proposition 5.** *Let $f$ be a mean function. The generalized Kulczynski similarity $GKU_f(\cdot,\cdot)$ satisfies triangle inequality if and only if $GKU_f(\cdot,\cdot) \equiv BB(\cdot,\cdot)$.*

**Proof.** The triangle inequality for distance measure $d(\mathcal{X},\mathcal{Y}) = 1 - s(\mathcal{X},\mathcal{Y})$ is stated as follows:

$$1 + s(\mathcal{X},\mathcal{Z}) \geq s(\mathcal{X},\mathcal{Y}) + s(\mathcal{Y},\mathcal{Z}) \text{ for all } \mathcal{X},\mathcal{Y},\mathcal{Z} \subseteq \mathcal{U}. \tag{21}$$

To prove the direct implication, we substitute an appropriate triple $\mathcal{X},\mathcal{Y},\mathcal{Z}$ in (21) for $s(\cdot,\cdot) = GKU_f(\cdot,\cdot)$. Let us consider two arbitrary positive reals $a < b$, and let us denote $r = \frac{b}{a} - 1 > 0$. Let us consider a finite set $\mathcal{Y}$ that is partitioned into two disjoint sets, $\mathcal{X}$ and $\mathcal{Z}$, such that $\mu(\mathcal{Z}) = r\mu(\mathcal{X})$, and let us denote $x = \mu(\mathcal{X})$. Substituting $\mathcal{X},\mathcal{Y},\mathcal{Z}$ in $s = GKU_f$, inequality (21) becomes the following:

$$1 + 0 \geq \frac{x}{f(x,(1+r)x)} + \frac{rx}{f(rx,(1+r)x)} = \frac{1}{f(1,1+r)} + \frac{r}{f(r,1+r)} \geq \frac{1}{1+r} + \frac{r}{1+r} = 1.$$

From this, we conclude that $f(1,1+r) = 1+r$.

It follows that $f(a,b) = af\left(1,\frac{b}{a}\right) = af(1,1+r) = a(1+r) = b = \max(a,b)$.

To prove the reverse implication, we slightly adapt Gilbert's proof of metricity of the Jaccard index from [4].

The distance is defined as follows:

$$d_{BB}(\mathcal{X},\mathcal{Y}) = 1 - BB(\mathcal{X},\mathcal{Y}) = 1 - \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{\max(\mu(\mathcal{X}),\mu(\mathcal{Y}))}. \tag{22}$$

The proof uses the notation of sets introduced in Figure 1. Let us denote $M = \max(\mu(\mathcal{X}),\mu(\mathcal{Y}),\mu(\mathcal{Z}))$. Then, it is not difficult to observe the following:

$$\begin{aligned}
&\frac{\max(\mu(\mathcal{X}')+\mu(\mathcal{Y}'')+\mu(\mathcal{Z}''),\mu(\mathcal{X}'')+\mu(\mathcal{Y}')+\mu(\mathcal{Z}''),\mu(\mathcal{X}'')+\mu(\mathcal{Y}'')+\mu(\mathcal{Z}'))}{M} = \\
&\frac{M-\mu(\mathcal{V})}{M} = 1 - \frac{\mu(\mathcal{V})}{M} \geq d_{BB}(\mathcal{X},\mathcal{Z}) \\
&d_{BB}(\mathcal{X},\mathcal{Y}) = \frac{\max(\mu(\mathcal{X}\setminus\mathcal{Y}),\mu(\mathcal{Y}\setminus\mathcal{X}))}{\max(\mu(\mathcal{X}),\mu(\mathcal{Y}))} \geq \frac{\max(\mu(\mathcal{X}')+\mu(\mathcal{Y}''),\mu(\mathcal{X}'')+\mu(\mathcal{Y}'))}{M} \\
&\text{By analogy } d_{BB}(\mathcal{Y},\mathcal{Z}) \geq \frac{\max(\mu(\mathcal{Y}')+\mu(\mathcal{Z}''),\mu(\mathcal{Y}'')+\mu(\mathcal{Z}'))}{M}
\end{aligned} \tag{23}$$

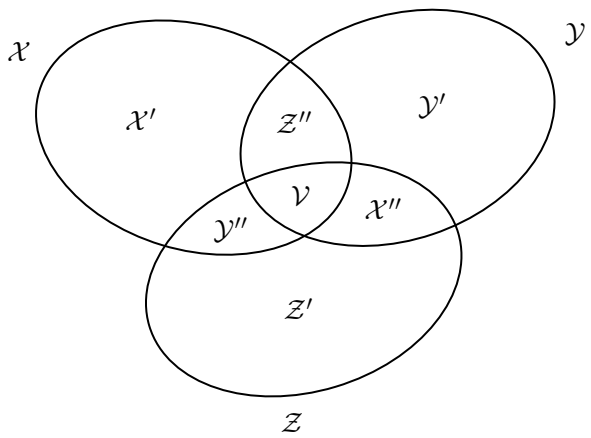


**Figure 1.** Three sets and their intersections.

Then, adding the last two inequalities of (23), applying $\max(a,b) + \max(c,d) = \max(a+c, a+d, b+c, b+d)$, and relating that to the first inequality of (23), we obtain the triangle inequality for $d_{BB}$. □

**Example 3.** *Let us consider the following set:*

$$\mathcal{Z} = \{1, 2, 4, 5, \ldots, 3n-2, 3n-1\}, \tag{24}$$

*together with sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1. Observe that $\mathcal{X}$, $\mathcal{Z}$ is a partition of $\mathcal{Y}$. It follows that*

$$\begin{aligned} GKU_f(\mathcal{X}, \mathcal{Z}) &= 0, \\ GKU_f(\mathcal{X}, \mathcal{Y}) &= \frac{n}{f(n, 3n+1)}, \\ GKU_f(\mathcal{Y}, \mathcal{Z}) &= \frac{2n+1}{f(n, 3n+1)}. \end{aligned}$$

*Now, for mean function $f \in \{AM, HM, GM, \min\}$, clearly $f(n, 3n+1) < 3n+1$. This shows the following:*

$$GKU_f(\mathcal{X}, \mathcal{Y}) + GKU_f(\mathcal{Y}, \mathcal{Z}) = \frac{3n+1}{f(n, 3n+1)} > 1 = 1 + GKU_f(\mathcal{X}, \mathcal{Z}),$$

*thus contradicting (21).*

## 6. Generalization of Jaccard Similarity

### *6.1. Definition and Basic Properties*

Following the introduction and argumentation of a set of axioms, the ratio model of normalized similarity known as the Tversky index is proposed in [16] according to the following equation with parameters $\alpha, \beta \geq 0$:

$$TV_{\alpha,\beta}(\mathcal{X}, \mathcal{Y}) = \frac{\mu(\mathcal{X} \cap \mathcal{Y})}{\mu(\mathcal{X} \cap \mathcal{Y}) + \alpha\mu(\mathcal{X} \setminus \mathcal{Y}) + \beta\mu(\mathcal{Y} \setminus \mathcal{X})}. \tag{25}$$

It is not difficult to observe that $TV_{\alpha,\beta}(\mathcal{X}, \mathcal{Y})$ is generally asymmetric if $\alpha \neq \beta$. Moreover, the symmetric version for $\alpha = \beta$ is a generalization of Jaccard similarity with some computational complexity as the standard Jaccard index.

**Proposition 6.** *Let us define the following:*

$$J_\alpha(\mathcal{X}, \mathcal{Y}) = TV_{\alpha,\alpha}(\mathcal{X}, \mathcal{Y}) = \frac{\mu(\mathcal{X} \cap \mathcal{Y})}{\mu(\mathcal{X} \cap \mathcal{Y}) + \alpha(\mu(\mathcal{X} \setminus \mathcal{Y}) + \mu(\mathcal{Y} \setminus \mathcal{X}))} \text{ for } \alpha > 0. \tag{26}$$

*Then,*

*i. $J_\alpha(\mathcal{X}, \mathcal{Y})$ is a generalization of Jaccard (5), Sørensen–Dice (6), and Sokal and Sneath (10) similarities.*

*ii. $J_\beta(\mathcal{X}, \mathcal{Y}) \geq J_\alpha(\mathcal{X}, \mathcal{Y})$ if $\alpha > \beta > 0$.*

**Proof.** Observe the following:

$$J_\alpha(\mathcal{X}, \mathcal{Y}) = \frac{a}{a + \alpha(b + c)}, \tag{27}$$

which clearly shows that

$$\begin{aligned} J(\mathcal{X}, \mathcal{Y}) &= J_1(\mathcal{X}, \mathcal{Y}), \\ SD(\mathcal{X}, \mathcal{Y}) &= J_{\frac{1}{2}}(\mathcal{X}, \mathcal{Y}), \\ SS(\mathcal{X}, \mathcal{Y}) &= J_2(\mathcal{X}, \mathcal{Y}), \end{aligned} \tag{28}$$

thus concluding the first point of the proposition.

If $\alpha > \beta > 0$, then the denominator of $J_\alpha(\mathcal{X}, \mathcal{Y})$ is greater than the denominator of $J_\beta(\mathcal{X}, \mathcal{Y})$, thus concluding the second point of the proposition. □

A corollary of Proposition 6 is the relative ordering of Jaccard (5), Sørensen–Dice (6), and Sokal and Sneath (10) similarities:

$$SS(\mathcal{X}, \mathcal{Y}) \leq J(\mathcal{X}, \mathcal{Y}) \leq SD(\mathcal{X}, \mathcal{Y}). \tag{29}$$

**Example 4.** *It is not difficult to observe that Inequality (29) is satisfied for sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1 by simple algebraic manipulation (first column of Table 1) and simple observation (second and third columns of Table 1).*

**Proposition 7.**

*i. For any mean function $f$ and $\alpha \geq 1$,*

$$J_\alpha(\mathcal{X}, \mathcal{Y}) \leq GKU_f(\mathcal{X}, \mathcal{Y}). \tag{30}$$

*ii. For any mean function $f$, such that $f(x, y) \leq AM(x, y)$,*

$$J_\alpha(\mathcal{X}, \mathcal{Y}) \leq GKU_f(\mathcal{X}, \mathcal{Y}), \tag{31}$$

*if and only if $\alpha \geq \frac{1}{2}$.*

**Proof.**

i. Inequality (30) is equivalent to the following:

$$a + b + c + (\alpha - 1)(b + c) \geq f(a + b, a + c).$$

This follows on from the fact that $f(a + b, a + c)$ is smaller than both $a + b$ and $a + c$, and $\alpha - 1 \geq 0$ and $a + b + c$ is greater than both $a + b$ and $a + c$.

ii. Let us assume that Inequality (31) holds. It follows that

$$a + \alpha(b + c) \geq f(a + b, a + c).$$

Substituting $b = c > 0$ and noticing that $f(a + b, a + b) = a + b$, we obtain $\alpha \geq \frac{1}{2}$.

Conversely, Inequality (31) is equivalent to the following:

$$\frac{(a + b) + (a + c)}{2} + (\alpha - \frac{1}{2})(b + c) = AM(a + b, a + c) + (\alpha - \frac{1}{2})(b + c) \geq f(a + b, a + c),$$

which follows on from $f(x, y) \leq AM(x, y)$ and $\alpha \geq \frac{1}{2}$. □

**Example 5.** *It is not difficult to observe that Inequality (30) is satisfied for sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1, $\alpha \in \{1, 2\}$ (producing similarity measures J and SS) and $f \in \{AM, HM, GM, \min, \max\}$ (producing similarity measures SD, KU, DKO, OV, and BB) by simple algebraic manipulation (first column of Table 1) and simple observation (second and third columns of Table 1).*

*Moreover, it is not difficult to observe that Inequality (31) is satisfied for sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1, $\alpha = \frac{1}{2}$ (producing similarity measure SD) and $f \in \{AM, HM, GM, \min\}$ (producing similarity measures SD, KU, DKO, and OV) by simple algebraic manipulation (first column of Table 1) and simple observation (second and third columns of Table 1).*

*6.2. Metric Properties*

The last part of this section is dedicated to a result concerning the metricity of generalized Jaccard similarity.

**Proposition 8.** *Generalized Jaccard similarity (26) satisfies triangle inequality if and only if* $\alpha \geq 1$.

**Proof.** For the direct implication, we consider a finite set $\mathcal{Y}$ partitioned into two disjointed subsets, $\mathcal{X}$ and $\mathcal{Z}$, of equal measures denoted by $n$. Applying inequality 21 for $s(\cdot) = J_\alpha(\cdot,\cdot)$, we obtain the following:

$$1 + 0 \geq \frac{n}{n+\alpha n} + \frac{n}{n+\alpha n} = \frac{2}{1+\alpha} \text{ so } \alpha \geq 1.$$

For the reverse implication, we adapt Gilbert's proof of metricity of Jaccard similarity from [4]. Referring to Figure 1, we define the following:

$$\begin{aligned} \mathcal{X}_1 &= \mathcal{X}' \cup \mathcal{X}'', \\ \mathcal{Y}_1 &= \mathcal{Y}' \cup \mathcal{Y}'', \\ \mathcal{Z}_1 &= \mathcal{Z}' \cup \mathcal{Z}'', \\ \mathcal{T} &= \mathcal{X} \cup \mathcal{Y} \cup \mathcal{Z}. \end{aligned} \tag{32}$$

Let us also introduce the following:

$$\begin{aligned} &\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z}) + \mu(\mathcal{X}\Delta\mathcal{Z}) = D, \\ &\mu(\mathcal{X}_1) + \mu(\mathcal{Y}_1) + \mu(\mathcal{Z}_1) = \frac{D}{2}. \end{aligned} \tag{33}$$

Now, observe that $\mu(\mathcal{X}\Delta\mathcal{Y})$ satisfies triangle inequality and has an upper bound of $\frac{D}{2}$, while $\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z})$ has a lower bound of $\frac{D}{2}$, as shown below.

$$\begin{aligned} &\mu(\mathcal{X}\Delta\mathcal{Y}) = \mu(\mathcal{X}_1) + \mu(\mathcal{Y}_1), \\ &\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z}) = \mu(\mathcal{X}_1) + 2\mu(\mathcal{Y}_1) + \mu(\mathcal{Z}_1) \geq \mu(\mathcal{X}_1) + \mu(\mathcal{Z}_1) = \mu(\mathcal{X}\Delta\mathcal{Z}), \\ &2\mu(\mathcal{X}\Delta\mathcal{Y}) \leq \mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{X}\Delta\mathcal{Z}) + \mu(\mathcal{Z}\Delta\mathcal{Y}) = D \text{ so } \mu(\mathcal{X}\Delta\mathcal{Y}) \leq \frac{D}{2}, \\ &2(\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z})) = (\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z})) + (\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z})) \geq, \\ &\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z}) + \mu(\mathcal{X}\Delta\mathcal{Z}) = D \text{ so:} \\ &\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z}) \geq \frac{D}{2}. \end{aligned} \tag{34}$$

Also, observe that if $\alpha \geq 1$, denoting $\theta = \alpha - 1 \geq 0$, we can rewrite generalized Jaccard similarity as follows:

$$J_\alpha(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X} \cap \mathcal{Y})}{\mu(\mathcal{X} \cup \mathcal{Y}) + \theta\mu(\mathcal{X}\Delta\mathcal{Y})}. \tag{35}$$

We obtain the following:

$$d_{J_\alpha}(\mathcal{X},\mathcal{Y}) = 1 - \frac{\mu(\mathcal{X}\cap\mathcal{Y})}{\mu(\mathcal{X}\cup\mathcal{Y}) + \theta\mu(\mathcal{X}\Delta\mathcal{Y})} \leq 1 - \frac{\mu(\mathcal{V})}{\mu(\mathcal{T}) + \theta\frac{D}{2}} =$$
$$\frac{\mu(\mathcal{X}_1) + \mu(\mathcal{Y}_1) + \mu(\mathcal{Z}_1) + \theta\frac{D}{2}}{\mu(\mathcal{T}) + \theta\frac{D}{2}} = \frac{D(1+\theta)}{2\mu(\mathcal{T}) + \theta D}$$
$$d_{J_\alpha}(\mathcal{X},\mathcal{Y}) = \frac{\mu(\mathcal{X}\cup\mathcal{Y}) - \mu(\mathcal{X}\cap\mathcal{Y}) + \theta\mu(\mathcal{X}\Delta\mathcal{Y})}{\mu(\mathcal{X}\cup\mathcal{Y}) + \theta\mu(\mathcal{X}\Delta\mathcal{Y})} \geq \frac{\mu(\mathcal{X}_1) + \mu(\mathcal{Y}_1) + \theta\mu(\mathcal{X}\Delta\mathcal{Y})}{\mu(\mathcal{T}) + \theta\frac{D}{2}}.$$

By analogy: $d_{J_\alpha}(\mathcal{Y},\mathcal{Z}) \geq \frac{\mu(\mathcal{Y}_1) + \mu(\mathcal{Z}_1) + \theta\mu(\mathcal{Y}\Delta\mathcal{Z})}{\mu(\mathcal{T}) + \theta\frac{D}{2}}$.

Adding last two inequalities:

$$d_{J_\alpha}(\mathcal{X},\mathcal{Y}) + d_{J_\alpha}(\mathcal{Y},\mathcal{Z}) \geq \frac{D + 2\theta(\mu(\mathcal{X}\Delta\mathcal{Y}) + \mu(\mathcal{Y}\Delta\mathcal{Z}))}{2\mu(\mathcal{T}) + \theta D} \geq \frac{D(1+\theta)}{2\mu(\mathcal{T}) + \theta D},$$

thus concluding the proof.

□

**Example 6.** *We check triangle inequality for $J_\alpha$ when $\alpha = \frac{1}{2}$ (i.e., for SD) and when $\alpha = 2$ (i.e., for SS). We consider sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1 and set $\mathcal{Z}$ from Example 3.*

*As $\mathcal{X}\cap\mathcal{Z} = \emptyset$, it follows that $SD(\mathcal{X},\mathcal{Z}) = SS(\mathcal{X},\mathcal{Z}) = 0$.*

$$SD(\mathcal{X},\mathcal{Y}) + SD(\mathcal{Y},\mathcal{Z}) = \frac{2n}{4n+1} + \frac{4n+2}{5n+2} > \frac{2n}{5n+2} + \frac{4n+2}{5n+2} = \frac{6n+2}{5n+2} > 1$$
$$= 1 + SD(\mathcal{X},\mathcal{Z}),$$

*thus contradicting (21), i.e., $J_{\frac{1}{2}} = SD$ does not satisfy triangle inequality.*

$$SS(\mathcal{X},\mathcal{Y}) + SS(\mathcal{Y},\mathcal{Z}) = \frac{n}{5n+2} + \frac{2n+1}{4n+1} < \frac{n}{4n+1} + \frac{2n+1}{4n+1} = \frac{3n+1}{4n+1} < 1$$
$$= 1 + SS(\mathcal{X},\mathcal{Z}),$$

*thus being consistent with (21), i.e., with $J_2 = SS$, satisfying triangle inequality.*

## 7. Generalization of Consonni and Todeschini Similarity

Consonni and Todeschini similarity [27] (see Equation (14)) can be generalized by replacing log with an appropriately chosen real $f$ function as follows:

$$GCT_f(\mathcal{X},\mathcal{Y}) = \frac{f(\mu(\mathcal{X}\cap\mathcal{Y}))}{f(\mu(\mathcal{X}\cup\mathcal{Y}))} = \frac{f(a)}{f(a+b+c)}. \tag{36}$$

Note that the computational overhead of $GCT_f$ compared to the specific similarity (14) is given solely by the computation of the function $f$.

For a proper similarity, we impose the condition $f(0) = 0$. Moreover, we shall assume in what follows that $f$ is defined only for non-negative values, and it is non-decreasing, which are natural assumptions. It follows immediately that $f$ is non-negative.

It is not difficult to observe that standard Consonni and Todeschini similarity is obtained for $f(x) = \log(1+x)$, i.e., $CT(\cdot,\cdot) = GCT_{\log(1+\cdot)}(\cdot,\cdot)$.

In what follows, we analyze the metric properties of generalized Consonni and Todeschini similarity. We start by formulating a necessary condition on $f$ to ensure that $GCT_f(\cdot,\cdot)$ satisfies triangle inequality.

**Proposition 9.** *If $GCT_f(\cdot,\cdot)$ satisfies triangle inequality, then $f$ is supermodular, i.e.,*

$$f(x+y+v)+f(v) \geq f(x+v)+f(y+v) \text{ for all } x,y,v \in \mathbb{R}^+. \tag{37}$$

**Proof.** Let us consider three sets, $\mathcal{X},\mathcal{Y}$, and $\mathcal{Z}=\mathcal{X}\cap\mathcal{Y}$, such that $\mu(\mathcal{X}\cap\mathcal{Y})=v$, $\mu(\mathcal{X}\setminus\mathcal{Y})$, and $\mu(\mathcal{X}\setminus\mathcal{Y})=y$. It follows that $\mathcal{X}=x+v$ and $\mathcal{Y}=y+v$. Applying Inequality (21), we obtain the following:

$$1+\frac{f(v)}{f(x+y+v)} \geq \frac{f(x+v)}{f(x+y+v)}+\frac{f(y+v)}{f(x+y+v)},$$

from which Inequality (37) follows immediately. □

It is not difficult to verify that function $\log(1+x)$ does not satisfy the supermodularity condition (37), clearly showing that standard Consonni and Todeschini similarity is not a metric.

**Example 7.** *We check triangle inequality for CT. Let us consider sets $\mathcal{X}$ and $\mathcal{Y}$ from Example 1 and set $\mathcal{Z}$ from Example 3.*

*As $\mathcal{X}\cap\mathcal{Z}=\emptyset$, it follows that $CT(\mathcal{X},\mathcal{Z})=0$.*

$$\begin{aligned} CT(\mathcal{X},\mathcal{Y})+CT(\mathcal{Y},\mathcal{Z}) &= \frac{\log(n+1)}{\log(3n+2)}+\frac{\log(2n+2)}{\log(3n+2)}=\frac{\log(2n^2+4n+2)}{\log(3n+2)} \\ &> \frac{\log(3n+2)}{\log(3n+2)}=1=1+CT(\mathcal{X},\mathcal{Z}), \end{aligned}$$

*thus contradicting (21), i.e., CT does not satisfy triangle inequality.*

Actually, it is not difficult to observe the following:

(i) If $f$ is strictly concave, then it is strictly submodular, i.e., it satisfies Inequality (37) in the opposite direction with strict inequality. So, according to Proposition 9, the corresponding generalized Consonni and Todeschini similarity is not a metric.

(ii) If $f$ is convex, then it is supermodular. In this case, the applicability of Proposition 10 can be investigated, as discussed in the following paragraphs.

Our next result is the formulation of a sufficient condition on $f$ that guarantees the corresponding generalized Consonni and Todeschini similarity (36) is a metric, i.e., it satisfies triangle inequality.

**Proposition 10.** *If $f$ is differentiable and supermodular and $\log\circ f$ is concave, then generalized Consonni and Todeschini similarity, defined by Equation (36), satisfies triangle inequality (21).*

**Proof.** With reference to Figure 1 and Equations (32), let us denote with lowercase letters $x,y,z,x',y',z',x'',y'',z'',x_1,y_1,z_1,t,v$, the measures of sets $\mathcal{X},\mathcal{Y},\mathcal{Z},\mathcal{X}',\mathcal{Y}',\mathcal{Z}',\mathcal{X}'',\mathcal{Y}'',\mathcal{Z}'',\mathcal{X}_1,\mathcal{Y}_1,\mathcal{Z}_1,\mathcal{T},\mathcal{V}$.

Substituting (36) into Equation (21) we obtain the following:

$$1+\frac{f(v+y'')}{f(t-y')} \geq \frac{f(v+z'')}{f(t-z')}+\frac{f(v+x'')}{f(t-x')}. \tag{38}$$

In order to prove Inequality (38), we need the following.

**Lemma 1.** *If $f$ is differentiable, then for each $\alpha\geq 0$, function $g$ defined as $g(x)=\frac{f(x)}{f(x+\alpha)}$, i.e., it is non-decreasing.*

**Proof.**

$$g'(x) = \frac{f'(x)f(x+\alpha) - f(x)f'(x+\alpha)}{f^2(x+\alpha)} = \frac{f(x)}{f(x+\alpha)}\left(\frac{f'(x)}{f(x)} - \frac{f'(x+\alpha)}{f(x+\alpha)}\right) =$$
$$\frac{f(x)}{f(x+\alpha)}\left((\log\circ f)'(x) - (\log\circ f)'(x+\alpha)\right) \geq 0$$
as $\log\circ f$ is concave and therefore $(\log\circ f)'$ is decreasing.

As $g'(x) \geq 0$, we conclude that $g$ is non-decreasing. □

Lemma 1 implies that if $\alpha \geq 0$ and $b \geq a \geq 0$, then the following inequality holds:

$$\frac{f(a)}{f(b)} \leq \frac{f(a+\alpha)}{f(b+\alpha)}. \tag{39}$$

Applying (39), we obtain the following:

$$\frac{f(v+z'')}{f(t-z')} \leq \frac{f(v+z''+z')}{f(t)} = \frac{f(v+z_1)}{f(t)}$$
$$\text{Similarly } \frac{f(v+x'')}{f(t-x')} \leq \frac{f(v+x_1)}{f(t)}$$
$$\text{Moreover } 1 + \frac{f(v+y'')}{f(t-y')} \geq 1 + \frac{f(v)}{f(t)} = 1 + \frac{f(v)}{f(v+x_1+y_1+z_1)}$$
Adding the first two inequalities, it is enough to show that:
$$f(v+x_1+y_1+z_1) + f(v) \geq f(v+x_1) + f(v+z_1).$$

But, this follows on from $f(v+x_1+y_1+z_1) \geq f(v+x_1+z_1)$ and from the supermdularity condition (37), thus concluding the proof. □

Examples of functions satisfying the requirements of Proposition 10 are as follows:

i. $x^p$ with $p \geq 1$.
ii. $e^x - 1$.
iii. $x\log(x)$ extended by continuity to 0 in $x = 0$.

Examples of functions that are not supermodular, i.e., the corresponding generalized Consonni and Todeschini similarity, fail to satisfy triangle inequality by Proposition 9 and are as follows:

i. $x^p$ with $p < 1$.
ii. $\log(x+1)$.

## 8. Conclusions

The results reported in this paper shed light on some theoretical properties of the most popular normalized similarity/distance measures for Boolean models that only depend on the characteristics present in compared objects. The new theoretical results covered the following three aspects: generalization of three classes of measures described as generalized Kulczynski, generalized Jaccard, and generalized Consonni and Todeschini measures; theoretical ordering of the similarity measures inside each class, as well as between classes; positive and negative results regarding the metric properties of the measures related to satisfying or not satisfying the triangle inequality axiom. In future work, we foresee at least two possibilities of continuing this theoretical research: by expanding the applicability of our methodology to other classes of similarity measures, possibly in the context of richer data models, like multisets, numerical, probabilistic, or fuzzy, as well as by investigating more abstract frameworks provided by lattice theory or other axiomatic approaches. On

the experimental side, it would be interesting to investigate the experimental comparison of similarity measures and the visualization of results on sample datasets of various sizes.

**Author Contributions:** All authors contributed equally to each section of the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** No new data were created or analyzed in this study.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Lesot, M.-J.; Rifqi, M.; Benhadda, H. Similarity measures for binary and numerical data: A survey. *Int. J. Knowl. Eng. Soft Data Paradig.* **2009**, *1*, 63–84. [CrossRef]
2. Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547–579.
3. Lewandowsky, M.; Winter, D. Distance between Sets. *Nature* **1971**, *234*, 34–35. [CrossRef]
4. Gilbert, G. Distance between Sets. *Nature* **1972**, *239*, 174. [CrossRef]
5. Kosub, S. A note on the triangle inequality for the Jaccard distance. *Pattern Recognit. Lett.* **2019**, *120*, 36–38. [CrossRef]
6. Gerasimou, G. Characterization of the Jaccard dissimilarity metric and a generalization. *Discret. Appl. Math.* **2024**, *355*, 57–61. [CrossRef]
7. Cheetham, A.H.; Hazel, J.E. Binary (presence-absence) similarity coefficients. *J. Paleontol.* **1969**, *43*, 1130–1136.
8. Hubálek, Z. Coefficients of Association and Similarity, Based on Binary (Presence-Absence) Data: An Evaluation. *Biol. Rev.* **1982**, *57*, 669–689. [CrossRef]
9. Brusco, M.; Cradit, J.D.; Steinley, D. A comparison of 71 binary similarity coefficients: The effect of base rates. *PLoS ONE* **2021**, *16*, e0247751. [CrossRef]
10. Rajski, C. A metric space of discrete probability distributions. *Inf. Control* **1961**, *4*, 371–377. [CrossRef]
11. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
12. Simovici, D.A. On Submodular and Supermodular Functions on Lattices and Related Structures. In Proceedings of the 2014 IEEE 44th International Symposium on Multiple-Valued Logic, Bremen, Germany, 19–21 May 2014; pp. 202–207.
13. Moulton, R.; Jiang, Y. Maximally Consistent Sampling and the Jaccard Index of Probability Distributions. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 347–356.
14. Berman, M.; Blaschko, M.B. Supermodular Locality Sensitive Hashes. *arXiv* **2018**, arXiv:1807.06686v1.
15. Deza, M.M.; Deza, E. *Encyclopedia of Distances*, 4th ed.; Springer: Berlin/Heidelberg, Germany, 2016.
16. Tversky, A. Features of Similarity. *Psychol. Rev.* **1977**, *84*, 327–352. [CrossRef]
17. Dice, L.R. Measures of the Amount of Ecologic Association Between Species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
18. Sørensen, T. A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons. *Biol. Skr. Dan. Vidensk. Selsk.* **1948**, *5*, 1–34.
19. Driver, E.S.; Kroeber, A.L. Quantitative Expression of Cultural Relationships. *Univ. Calif. Publ. Am. Archaeol. Ethnol.* **1932**, *31*, 211–256.
20. Ochiai, A. Zoogeographical Studies on the Soleoid Fishes Found in Japan and its Neighbouring Regions-III. *Nippon. Suisan Gakkaishi* **1957**, *22*, 522–525. [CrossRef]
21. Sorgenfrei, T. Molluscan assemblages from the marine middle Miocene of South Jutland and their environments. *Den. Geol. Undersoegelse Ser. II* **1958**, *79*, 356–503.
22. Sokal, R.R.; Sneath, P.H.A. *Principles of Numerical Taxonomy*; W. H. Freeman and Co.: San Francisco, CA, USA; London, UK, 1963.
23. Kulczynski, S. Die Pflanzenassoziationen der Pieninen. *Bull. Int. L'AcadéMie Pol. Sci. Lett. Classe Sci. Math. Nat. B (Sci. Nat.)* **1927**, 57–203.
24. McGill, M. An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. *ERIC Inst. Educ. Sci.* **1979**, ED188587.
25. Simpson, G.G. Notes on the Measurement of Faunal Resemblance. *Am. J. Sci.* **1960**, *258-A*, 300–311.
26. Braun-Blanquet, J. Zur Wertung der Gesellschaftstreue in der Pflanzensoziologie. Vierteljahrsschr. *Naturf. Ges. Zürich* **1925**, *70*, 12–149.

27. Todeschini, R.; Consonni, V.; Xiang, H.; Holliday, J.; Buscema, M.; Willett, P. Similarity Coefficients for Binary Chemoinformatics Data: Overview and Extended Comparison Using Simulated and Real Data Sets. *J. Chem. Inf. Model.* **2012**, *52*, 2884–2901. [CrossRef]
28. Beliakov, G.; Bustince Sola, H.; Calvo Sánchez, T. Classical Averaging Functions. In: A Practical Guide to Averaging Functions. *Stud. Fuzziness Soft Comput.* **2016**, *329*, 55–99.