Discrete Applied Mathematics 355 (2024) 57–61

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

Note

Characterization of the Jaccard dissimilarity metric and a generalization☆

Georgios Gerasimou

University of Glasgow, United Kingdom

ARTICLE INFO

Article history:
Received 24 April 2023
Received in revised form 27 March 2024
Accepted 28 April 2024
Available online 9 May 2024

Keywords:
Jaccard metric
Dissimilarity
Axioms
Characterization
Quasi-metric.

ABSTRACT

The Jaccard dissimilarity metric identifies the distance between two finite sets by the number of their unique elements as a proportion of their joint cardinality. This note gives an elementary characterization of the Jaccard metric by means of three simple axioms. Relaxing the most substantial one allows for a general family of new dissimilarity quasi-metrics to emerge that encompasses Jaccard's metric.

1. Introduction

For two finite sets A and B that are not both empty, the Jaccard metric [7–12] defines the distance between them by

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

This simple and intuitive formula is often interpreted as reflecting the dissimilarity between A and B. It is frequently applied in such diverse fields as ecology, genetics, meteorology, information theory, operations research and recommender systems [1,2,4,6,13,15,16].

While axiomatic characterizations of other dissimilarity measures in different domains are known [3,5,14], it appears that this is not the case for (1). We provide an elementary such characterization by means of three simple axioms. These do not include the triangle inequality although, of course, they imply it.

Our analysis highlights the following two properties that are implicit in (1): (i) dissimilarity increases by a constant quantity when any element belonging to two distinct sets is removed from one but not both sets (**A3**; Section 2); (ii) the cumulative dissimilarity between each of two disjoint sets and their union is the maximum possible (**A6**; Section 3). Relaxing these two requirements allows for a general and, as we show, novel family of dissimilarity quasi-metrics to emerge, which includes Jaccard's metric as a special case. Our analysis, finally, helps shed some new light on three distinct and pre-existing such quasi-metrics.

☆ I thank a referee for constructive suggestions that strengthened the characterization and shortened its proof.
E-mail address: Georgios.Gerasimou@glasgow.ac.uk.

https://doi.org/10.1016/j.dam.2024.04.022

2. Axioms for the Jaccard metric

We assume an unstructured finite set X and consider a mapping $R : 2^X \times 2^X \to \mathbb{R}$ that is well-defined everywhere except at $(\emptyset, \emptyset)$ (it will be without loss to assume throughout that it takes the value of zero at that point). **A1**-**A3** below are some properties that R could have if $R(A, B)$ were to be interpreted as reflecting the dissimilarity between sets A and B in this domain:

A1. $R(A, B) = R(B, A)$.

A2. $R(A, B) = 0 \iff A = B$.

A3. $A \not\ni x \in B \implies R(A, B) - R(A \cup \{x\}, B) = \dfrac{1}{|A \cup B|}$.

A1 and **A2** are the standard metric symmetry and identity properties. **A3** requires constant marginal sensitivity of R to the gradual removal from one of the two sets of elements that belong to both sets. In addition, it requires such constant marginal sensitivity to be reciprocally dependent on the joint cardinality of the two sets, which of course remains fixed during this gradual removal process.

Proposition 1. *R satisfies **A1**, **A2**, **A3** if and only if $R \equiv J$. Moreover, these axioms are independent.*

Proof. It is immediate that these axioms are implied by J. For the converse implication, let $R : 2^X \times 2^X \to \mathbb{R}$ satisfy **A1**, **A2**, **A3**. Consider $A, B \subseteq X$, with A, B not both empty. It will be shown that, regardless of whether $A \subseteq B$, $B \subseteq A$, or otherwise, we have

$$\begin{aligned} R(A, B) &= \frac{|A \setminus B|}{|A \cup B|} + \frac{|B \setminus A|}{|A \cup B|} \\ &\equiv \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \\ &\equiv J(A, B) \end{aligned} \tag{2}$$

Suppose first that there is $x \in B \setminus A$. By **A3**, $R(A, B) = R(A \cup \{x\}, B) + \frac{1}{|A \cup B|}$. Applying **A3** iteratively over all elements of $B \setminus A$ yields

$$R(A, B) = R(A \cup B, B) + \frac{|B \setminus A|}{|A \cup B|}. \tag{3}$$

Now, either there is $x \in A \setminus B$ or $A \subseteq B$. Consider the former case first. By **A3**, $R(B, A \cup B) = R(B \cup \{x\}, A \cup B) + \frac{1}{|A \cup B|}$ for such x. Applying **A3** iteratively over all elements of $A \setminus B$ gives

$$\begin{aligned} R(A \cup B, B) &= R(B, A \cup B) \\ &= R(A \cup B, A \cup B) + \frac{|A \setminus B|}{|A \cup B|} \\ &= \frac{|A \setminus B|}{|A \cup B|}, \end{aligned} \tag{4}$$

with the first step following from **A1** and the last from **A2**. If $A \subseteq B$ holds instead, then

$$\begin{aligned} R(A \cup B, B) &= R(B, B) \\ &= 0, \end{aligned} \tag{5}$$

by **A2**.

Now suppose there is no $x \in B \setminus A$, so that $B \subseteq A$. This implies

$$R(A, B) = R(A \cup B, B). \tag{6}$$

If $B \subset A$, then (4) and (6) imply

$$R(A, B) = \frac{|A \setminus B|}{|A \cup B|}. \tag{7}$$

If $A = B$ instead, then (5) and (6) imply $R(A, B) = 0$. Thus, (2) holds in all cases.

To verify that **A1**, **A2** and **A3** are independent we show that there is a set X such that, for every $i, j \in \{1, 2, 3\}$, there exists $R_j : 2^X \times 2^X \to \mathbb{R}$ that satisfies each axiom **A***i* except **A***j*. To this end, it suffices to consider the domains $X := \{a, b\}$ and $2^X \equiv \big\{\emptyset, \{a\}, \{b\}, \{a, b\}\big\}$. In each of the examples below that are presented in tabular form, the row and column entries, respectively, represent the first and second arguments of the corresponding R_i.

R_1:

	$\emptyset$	$\{a\}$	$\{b\}$	$\{a, b\}$
$\emptyset$	0	1	1	1
$\{a\}$	$\frac{1}{3}$	0	1	$\frac{1}{2}$
$\{b\}$	$\frac{1}{3}$	1	0	$\frac{1}{2}$
$\{a, b\}$	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{1}{2}$	0

R_2:

	$\emptyset$	$\{a\}$	$\{b\}$	$\{a, b\}$
$\emptyset$	0	1	1	1
$\{a\}$	1	0	1	$\frac{1}{2}$
$\{b\}$	1	1	0	$\frac{1}{2}$
$\{a, b\}$	1	$\frac{1}{2}$	$\frac{1}{2}$	1

R_3:

	$\emptyset$	$\{a\}$	$\{b\}$	$\{a, b\}$
$\emptyset$	0	1	1	1
$\{a\}$	1	0	1	$\frac{1}{3}$
$\{b\}$	1	1	0	$\frac{1}{3}$
$\{a, b\}$	1	$\frac{1}{3}$	$\frac{1}{3}$	0

Remark. The proof only uses the part of **A2** that requires $R(A, A) = 0$ for all A.

3. A general class of dissimilarity quasi-metrics

The constant marginal sensitivity axiom **A3** may be unappealing in human dissimilarity judgments within collections of pairs of sets such as the one depicted in Fig. 1. There, $B \supset A_1 \supset A_2 \supset \cdots \supset A_9 \supset A_{10}$; $|B| - |A_1| = |A_j| - |A_{j+1}| = 1$ for $j \leq 9$; and $J(A_1, B) - J(A_2, B) = J(A_9, B) - J(A_{10}, B) = \frac{1}{12}$. But since A_1 and A_2 are less Jaccard-dissimilar than A_9 and A_{10}, one might have expected the dissimilarity difference between A_1, B and A_2, B to be perceived as being smaller than that between A_9, B and A_{10}, B.

A simple generalization of (1) that allows for such potentially *increasing* marginal sensitivity to the gradual removal of common elements is achievable by letting the mapping $J^w : 2^X \times 2^X \to \mathbb{R}$ be defined by

$$J^w(A, B) := \frac{|A \cup B|^\alpha - |A \cap B|^\beta}{|A \cup B|^\alpha} = 1 - \frac{|A \cap B|^\beta}{|A \cup B|^\alpha}, \tag{8}$$

for some $1 \geq \alpha \geq \beta > 0$, with $\alpha \geq \beta$ sufficing for J^w to be non-negative.

Moreover, it is easily seen that, whenever $\alpha = \beta \leq 1$, J^w obeys the first two of the next three additional conditions that are satisfied by J (and hence are implied by **A1**-**A3**):

A4. $R(A, B) \in [0, 1]$.

A5. $R(A, B) = 1 \iff A \cap B = \emptyset$.

A6. $A \cap B = \emptyset \implies R(A, A \cup B) + R(B, A \cup B) = 1$.

A4 is a boundedness property with an obvious (dis)similarity meaning. In conjunction with the normalization imposed by **A4**, **A5** can be intuitively thought of as requiring that any two sets be maximally dissimilar if and only if they have nothing in common, and **A6** as demanding that there be maximum cumulative dissimilarity between each of two disjoint sets and their union.

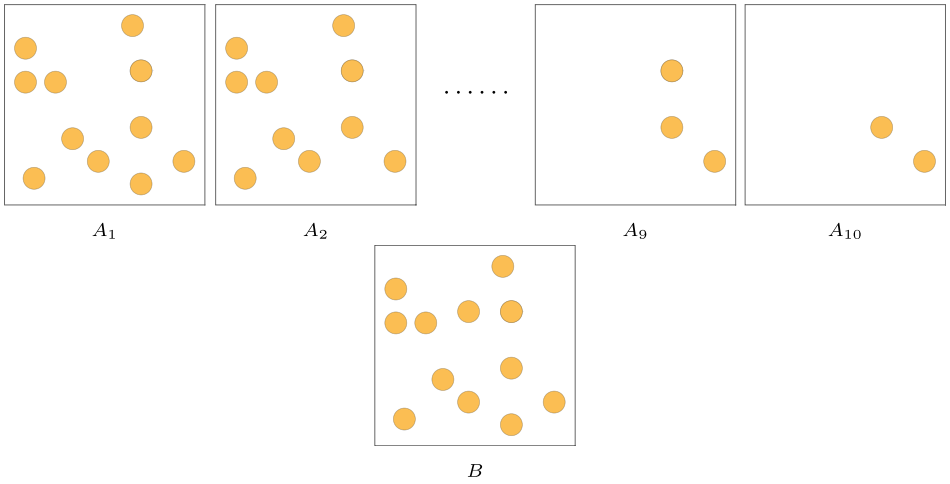

Fig. 1. J shows constant marginal sensitivity to the removal of jointly owned elements. Here, $J(A_1, B) - J(A_2, B) = J(A_9, B) - J(A_{10}, B)$.

Both **A3** and **A6** are violated by J^w whenever $\alpha \neq 1 \neq \beta$. However, J^w obeys the following *mid-point convexity* generalization of **A3** and *sub-additivity* generalization of **A6** if $\alpha = \beta \leq 1$:

A3′. $A \not\ni x, y \in B \implies R(A, B) - R(A \cup \{x\}, B) \geq R(A \cup \{x\}, B) - R(A \cup \{x, y\}, B)$.

A6′. $A \cap B = \emptyset \implies R(A, A \cup B) + R(B, A \cup B) \leq 1$.

We can now proceed with the following statement.

Proposition 2. *$J^w \equiv J \iff \alpha = \beta = 1$. Moreover, if $\alpha = \beta \leq 1$, then J^w implies* **A1**, **A2**, **A3′**, **A4**, **A5**, **A6′**. *The converse of this statement is false.*

Proof. The first part is straightforward. Regarding the second part, we note that R_3 in the proof of Proposition 1 satisfies the six axioms in the statement of Proposition 2 and coincides with the specific J^w where $\alpha = \beta = \frac{\ln(3)-\ln(2)}{\ln(2)} \approx 0.5849$. On the slightly richer set $X := \{a, b, c\}$, however, we provide below an example of some R that satisfies these axioms but is not a J^w function, thereby also proving the non-obvious part of Proposition 2.

Indeed, consider $X := \{a, b, c\}$, $2^X \equiv \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$ and let $\widehat{R} : 2^X \times 2^X \to \mathbb{R}$ be defined as follows:

$$
\begin{aligned}
\widehat{R}(A, B) &= 0 &&\iff A = B,\\
\widehat{R}(A, B) &= 1 &&\iff A \cap B = \emptyset,\\
\widehat{R}(A, B) &= \widehat{R}(B, A) && \forall A, B \in 2^X,\\
\widehat{R}(\{x, y\}, \{y, z\}) &= \frac{1}{2} && \forall x, y, z \in X : x \neq y \neq z \neq x,\\
\widehat{R}(\{x\}, \{x, y\}) &= \frac{1}{3} && \forall x, y \in X : x \neq y,\\
\widehat{R}(\{x\}, X) &= \frac{1}{4} && \forall x \in X :\\
\widehat{R}(\{x, y\}, X) &= \frac{1}{8} && \forall x, y \in X : x \neq y.
\end{aligned}
$$

We leave it as an exercise for the reader to verify that $\widehat{R}$ violates **A3** and **A6** but satisfies all axioms appearing in the statement of Proposition 2. That there is no $\alpha \in (0, 1)$ such that $\widehat{R} \equiv J^w$ under this α can be seen, for example, by observing that $\widehat{R}(\{a\}, \{a, b\}) = \frac{1}{3}$ is uniquely compatible with J^w under the α associated with function R_3 above, whereas $\widehat{R}(\{a, b\}, \{b, c\}) = \frac{1}{2}$ is so under the distinct $\alpha' = \frac{\ln(2)}{\ln(3)}$. Yet, J^w requires the value of this scalar to be invariant across all pairs.

One may wonder whether, in addition to **A3** and **A6**, J^w generally violates the triangle inequality too when $\alpha = \beta < 1$:

A7. $R(A, B) + R(B, C) \geq R(A, C)$.

The positive answer to this question can be confirmed, for example, at $\alpha = \beta := 0.5, A := \{1, 2, 3, 4\}, B := \{3, 4, 5, 6, 7\}$ and $C := \{5, 6, 7, 8\}$. This fact and Proposition 2 now imply that J^w with $\alpha = \beta < 1$ becomes a *quasi-metric* that features

Table 1
Axioms satisfied or violated by the different dissimilarity functions.

	A1	A2	A3	A3′	A4	A5	A6	A6′	A7
J	✓	✓	✓	✓	✓	✓	✓	✓	✓
J^w	✓	✓	×	✓	✓	✓	×	✓	×
SD	✓	✓	×	✓	✓	✓	×	✓	×
SC	✓	✓	×	✓	✓	✓	×	✓	×
$Overlap$	✓	✓	×	×	✓	✓	×	✓	×

increasing marginal sensitivity to the gradual removal of common elements. We can use the axioms introduced above to compare the structure of this class of dissimilarity quasi-metrics to other such functions in the existing literature.

To this end, we recall the Sørensen–Dice (SD), Salton cosine (SC) and $Overlap$ dissimilarity functions (see [15] and references therein), which are defined by

$$SD(A,B) := 1 - \frac{2|A\cap B|}{|A|+|B|} \tag{9}$$

$$SC(A,B) := 1 - \frac{|A\cap B|}{\sqrt{|A|\times|B|}} \tag{10}$$

$$Overlap(A,B) := 1 - \frac{|A\cap B|}{\min\{|A|,|B|\}} \tag{11}$$

As can be easily checked, all three satisfy **A1**, **A2**, **A4**, **A5** and violate **A3**, **A6**, **A7**. Furthermore, SD and SC satisfy **A3′**, whereas $Overlap$ violates this axiom (e.g. at $A=\{a,b\}$ and $B=\{b,c,d\}$ with $x=c$, $y=d$). Finally, they all satisfy **A6′**, and $Overlap$ does so trivially (the relevant sum is either 0 or 1). This information is summarized in Table 1.

Despite SD, SC and J^w sharing the same features as far as these 9 axioms are concerned, it is straightforward to see that the values of different pairs of sets (A,B), (A',B') under each of the terms that appear with a negative sign on the right hand sides of (9) and (10) are generally distinct from those under the corresponding term in the single-parameter specification of (8) with a fixed $\alpha\in(0,1)$. Thus, (8) and (9)–(11) are non-nested classes of dissimilarity quasi-metrics, with the one defined by (8) when $\alpha=\beta\leq 1$ apparently being a novel addition to that category. A complete characterization of this general class is left as an open problem.

Data availability

No data was used for the research described in the article.

References

[1] S. Azaele, R. Muneepeerakul, A. Maritan, I. Rodriguez-Iturbe, Predicting spatial similarity of freshwater fish biodiversity, Proc. Natl. Acad. Sci. USA 106 (17) (2009) 7058–7062.
[2] S. Bag, S.K. Kumar, M.K. Tiwari, An efficient recommendation generation using relevant Jaccard similarity, Inf. Sci. 483 (2019) 53–64.
[3] C. Bertoluzza, M. Di Bacco, V. Doldi, An axiomatic characterization of the measures of similarity, Sankhyā 66 (3) (2004) 474–486.
[4] M. Besta, R. Kanakagiri, H. Mustafa, M. Karasikov, G. Rätsch, T. Hoefler, E. Solomonik, Communication-efficient Jaccard similarity for high-performance distributed genome comparisons, IEEE Trans. Parallel Distrib. Syst. 2020 (2020) 1122–1132.
[5] D. Bouyssou, T. Marchant, M. Pirlot, Axiomatic characterization of the χ^2 dissimilarity measure, Aequat. Math. 96 (2022) 307–323.
[6] G.F. Gilbert, Finley's tornado predictions, Amer. Meteor. J. 1 (5) (1884) 166–172.
[7] A. Grygorian, I.E. Iacob, A concise proof of the triangle inequality for the Jaccard distance, Coll. Math. J. 49 (5) (2018) 363–365.
[8] Lipkus A. H., A proof of the triangle inequality for the Tanimoto distance, J. Math. Chem. 26 (1999) 263–265.
[9] P. Jaccard, Étude comparative de la distribution florale dans une portion des alpes et des jura, Bull. Soc. Vaudoise Sci. Nat. 37 (142) (1901) 547–579.
[10] S. Kosub, A note on the triangle inequality for the Jaccard distance, Pattern Recognit. Lett. 120 (2019) 36–38.
[11] M. Levandowsky, D. Winter, Distance between sets, Nature 234 (5) (1971) 34–35.
[12] D.J. Rogers, T.T. Tanimoto, A computer program for classifying plants, Science 132 (3434) (1960) 1115–1118.
[13] J.T. Schaefer, The critical success index as an indicator of warning skill, Weather Forecast. 5 (4) (1990) 570–575.
[14] A. Tversky, Features of similarity, Psychol. Rev. 84 (4) (1977) 327–352.
[15] V. Verma, R.K. Aggarwal, A comparative analysis of similarity measures akin to the Jaccard index in collaborative recommendations: Empirical and theoretical perspective, Soc. Netw. Anal. Min. 10 (2020) 43.
[16] I.E. Vorontsov, I.V. Kulakovskiy, V.J. Makeev, Jaccard index based similarity measure to compare transcription factor binding site models, Algorithms Mol. Biol. 8 (2013) 23.