Jaccard

Bjørn Kjos-Hanssen

December 12, 2024

This project contains excerpt from the paper *Interpolating between the Jaccard distance and an analogue of the normalized information distance*. That paper is unique in that a majority of the results were formalized at the time of publication. Therefore it is especially suitable for a Lean blueprint project.

Abstract. Jiménez, Becerra, and Gelbukh (2013) defined a family of "symmetric Tversky ratio models" $S_{\alpha,\beta}$, $0 \le \alpha \le 1$, $\beta > 0$. Each function $D_{\alpha,\beta} = 1 - S_{\alpha,\beta}$ is a semimetric on the powerset of a given finite set.

We show that $D_{\alpha,\beta}$ is a metric if and only if $0 \le \alpha \le \frac{1}{2}$ and $\beta \ge 1/(1-\alpha)$. This result is formally verified in the Lean proof assistant.

The extreme points of this parametrized space of metrics are $\mathcal{V}_1 = D_{1/2,2}$, the Jaccard distance, and $\mathcal{V}_\infty = D_{0,1}$, an analogue of the normalized information distance of M. Li, Chen, X. Li, Ma, and Vitányi (2004).

As a second interpolation, in general we also show that $\mathcal{V}_p$ is a metric, $1 \le p \le \infty$, where

$$\Delta_p(A,B) = (|B \setminus A|^p + |A \setminus B|^p)^{1/p},$$

$$\mathcal{V}_p(A,B) = \frac{\Delta_p(A,B)}{|A \cap B| + \Delta_p(A,B)}.$$

0.1 Introduction

Distance measures (metrics), are used in a wide variety of scientific contexts. In bioinformatics, M. Li, Badger, Chen, Kwong, and Kearney [13] introduced an information-based sequence distance. In an information-theoretical setting, M. Li, Chen, X. Li, Ma and Vitányi [14] rejected the distance of [13] in favor of a *normalized information distance* (NID). The Encyclopedia of Distances [3] describes the NID on page 205 out of 583, as

$$\frac{\max\{K(x \mid y^*), K(y \mid x^*)\}}{\max\{K(x), K(y)\}}$$

where $K(x \mid y^*)$ is the Kolmogorov complexity of x given a shortest program y^* to compute y. It is equivalent to be given y itself in hard-coded form:

$$\frac{\max\{K(x \mid y), K(y \mid x)\}}{\max\{K(x), K(y)\}}$$

Another formulation (see [14, page 8]) is

$$\frac{K(x,y) - \min\{K(x), K(y)\}}{\max\{K(x), K(y)\}}.$$

The fact that the NID is in a sense a normalized metric is proved in [14]. Then in 2017, while studying malware detection, Raff and Nicholas [15] suggested Lempel–Ziv Jaccard distance (LZJD) as a practical alternative to NID. As we shall see, this is a metric. In a way this constitutes a full circle: the distance in [13] is itself essentially a Jaccard distance, and the LZJD is related to it as Lempel–Ziv complexity is to Kolmogorov complexity. In the present paper we aim to shed light on this back-and-forth by showing that the NID and Jaccard distances constitute the endpoints of a parametrized family of metrics.

Reference	Jaccard notation	NID notation
[13]	d	
[14]	d_s	d
[10]	D	D'
[15]	LZJD	NCD

Table 1: Overview of notation used in the literature. (It seems that authors use simple names for their favored notions.)

For comparison, the Jaccard distance between two sets X and Y, and our analogue of the NID, are as follows:

$$J_1(X,Y) = \frac{|X \setminus Y| + |Y \setminus X|}{|X \cup Y|} = 1 - \frac{|X \cap Y|}{|X \cup Y|} \tag{1}$$

$$J_\infty(X,Y) = \frac{\max\{|X \setminus Y|, |Y \setminus X|\}}{\max\{|X|, |Y|\}} \tag{2}$$

Our main result Theorem 20 shows which interpolations between these two are metrics. The way we arrived at J_∞ as an analogue of NID is via Lempel–Ziv complexity. While there are several variants [12, 19, 20], the LZ 1978 complexity [20] of a sequence is the cardinality of a certain set, the dictionary.

Definition 1. Let $\mathrm{LZSet}(A)$ be the Lempel–Ziv dictionary for a sequence A. We define LZ–Jaccard distance LZJD by

$$\mathrm{LZJD}(A,B) = 1 - \frac{|\mathrm{LZSet}(A) \cap \mathrm{LZSet}(B)|}{|\mathrm{LZSet}(A) \cup \mathrm{LZSet}(B)|}.$$

It is shown in [13, Theorem 1] that the triangle inequality holds for a function which they call an information-based sequence distance. Later papers give it the notation d_s in [14, Definition V.1], and call their normalized information distance d. Raff and Nicholas [15] introduced the LZJD and did not discuss the appearance of d_s in [14, Definition V.1], even though they do cite [14] (but not [13]).

Kraskov et al. [11, 10] use D and D' for continuous analogues of d_s and d in [14] (which they cite). The *Encyclopedia* calls it the normalized information metric,

$$\frac{H(X \mid Y) + H(X \mid Y)}{H(X,Y)} = 1 - \frac{I(X;Y)}{H(X,Y)}$$

or Rajski distance [16].

This d_s was called d by [13] — see Table 1. Conversely, [14, near Definition V.1] mentions mutual information.

Remark 2. Ridgway [4] observed that the entropy-based distance D is essentially a Jaccard distance. No explanation was given, but we attempt one as follows. Suppose X_1, X_2, X_3, X_4 are iid Bernoulli($p = 1/2$) random variables, Y is the random vector (X_1, X_2, X_3) and Z is (X_2, X_3, X_4). Then Y and Z have two bits of mutual information $I(Y,Z) = 2$. They have an entropy $H(Y) = H(Z) = 3$ of three bits. Thus the relationship $H(Y,Z) = H(Y) + H(Z) - I(Y,Z)$ becomes a Venn diagram relationship $|\{X_1, X_2, X_3, X_4\}| = |\{X_1, X_2, X_3\}| + |\{X_2, X_3, X_4\}| - |\{X_2, X_3\}|$. The relationship to Jaccard distance may not have been well known, as it is not mentioned in [10, 2, 13, 1].

A more general setting is that of STRM (Symmetric Tversky Ratio Models), Definition 17. These are variants of the Tversky index (Definition 14) proposed in [7].

0.1.1 Generalities about metrics

Definition 3. Let $\mathcal{X}$ be a set. A *metric* on $\mathcal{X}$ is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that

1. $d(x, y) \geq 0$,
2. $d(x, y) = 0$ if and only if $x = y$,
3. $d(x, y) = d(y, x)$ (symmetry),
4. $d(x, y) \leq d(x, z) + d(z, y)$ (the triangle inequality)

for all $x, y, z \in \mathcal{X}$. If d satisfies Item 1, Item 2, Item 3 but not necessarily Item 4 then d is called a *semimetric*.

A basic exercise in Definition 3 that we will make use of is Theorem 4.

Theorem 4. *If d_1 and d_2 are metrics and a, b are nonnegative constants, not both zero, then $ad_1 + bd_2$ is a metric.*

Proof. Item 1 is immediate from Item 1 for d_1 and d_2.

Item 2: Assume $ad_1(x, y) + bd_2(x, y) = 0$. Then $ad_1(x, y) = 0$ and $bd_2(x, y) = 0$. Since a, b are not both 0, we may assume $a > 0$. Then $d_1(x, y) = 0$ and hence $x = y$ by Item 2 for d_1.

Item 3: We have $ad_1(x, y) + bd_2(x, y) = ad_1(y, x) + bd_2(y, x)$ by Item 3 for d_1 and d_2.

Item 4: By Item 4 for d_1 and d_2 we have

$$\begin{aligned} ad_1(x, y) + bd_2(x, y) &\leq a(d_1(x, z) + d_1(z, y)) + b(d_2(x, z) + d_2(z, y)) \\ &= (ad_1(x, z) + bd_2(x, z)) + (ad_2(z, y) + bd_2(z, y)). \square \end{aligned}$$

Lemma 5. *Let $d(x, y)$ be a metric and let $a(x, y)$ be a nonnegative symmetric function. If $a(x, z) \leq a(x, y) + d(y, z)$ for all x, y, z, then $d'(x, y) = \frac{d(x,y)}{a(x,y)+d(x,y)}$, with $d'(x, y) = 0$ if $d(x, y) = 0$, is a metric.*

Proof. As a piece of notation, let us write $d_{xy} = d(x, y)$ and $a_{xy} = a(x, y)$. As observated by [17], in order to show

$$\frac{d_{xy}}{a_{xy} + d_{xy}} \leq \frac{d_{xz}}{a_{xz} + d_{xz}} + \frac{d_{yz}}{a_{yz} + d_{yz}},$$

it suffices to show the following pair of inequalities:

$$\frac{d_{xy}}{a_{xy} + d_{xy}} \leq \frac{d_{xz}+d_{yz}}{a_{xy}+d_{xz}+d_{yz}} \tag{3}$$

$$\frac{d_{xz}+d_{yz}}{a_{xy}+d_{xz}+d_{yz}} \leq \frac{d_{xz}}{a_{xz} + d_{xz}} + \frac{d_{yz}}{a_{yz} + d_{yz}} \tag{4}$$

Here (3) follows from d being a metric, i.e., $d_{xy} \leq d_{xz} + d_{yz}$, since

$$c \geq 0 < a \leq b \implies \frac{a}{a + c} \leq \frac{b}{b + c}.$$

Next, (4) would follow from $a_{xy} + d_{yz} \geq a_{xz}$ and $a_{xy} + d_{xz} \geq a_{yz}$. By symmetry between x and y and since $a_{xy} = a_{yx}$ by assumption, it suffices to prove the first of these, $a_{xy} + d_{yz} \geq a_{xz}$, which holds by assumption. $\square$

0.1.2 Metrics on a family of finite sets

Lemma 6. *For sets A, B, C, we have $|A \setminus B| \leq |A \setminus C| + |C \setminus B|$.*

Proof. We have $A \setminus B \subseteq (A \setminus C) \cup (C \setminus B)$. Therefore, the result follows from the union bound for cardinality. □

Lemma 7. *Let $f(A, B) = |A \setminus B| + |B \setminus A|$. Then f is a metric.*

Proof. The most nontrivial part is to prove the triangle inequality,

$$|A \setminus B| + |B \setminus A| \leq |A \setminus C| + |C \setminus A| + |C \setminus B| + |B \setminus C|.$$

By the "rotation identity" $|A \setminus C| + |C \setminus B| + |B \setminus A| = |A \setminus B| + |B \setminus C| + |C \setminus A|$, this is equivalent to

$$2(|A \setminus B| + |B \setminus A|) \leq 2(|A \setminus C| + |C \setminus B| + |B \setminus A|),$$

which is immediate from Lemma 6. □

Lemma 8. *Let $f(A, B) = \max\{|A \setminus B|, |B \setminus A|\}$. Then f is a metric.*

Proof. For the triangle inequality, we need to show

$$\max\{|A \setminus B|, |B \setminus A|\} \leq \max\{|A \setminus C|, |C \setminus A|\} + \max\{|C \setminus B|, |B \setminus C|\}.$$

By symmetry we may assume that $\max\{|A \setminus B|, |B \setminus A|\} = |A \setminus B|$. Then, the result is immediate from Lemma 6. □

For a real number α, we write $\overline{\alpha} = 1 - \alpha$. For finite sets X, Y we define

$$\tilde{m}(X, Y) = \min\{|X \setminus Y|, |Y \setminus X|\},$$

$$\tilde{M}(X, Y) = \max\{|X \setminus Y|, |Y \setminus X|\}.$$

Lemma 9. *Let $\delta := \alpha\tilde{m} + \overline{\alpha}\tilde{M}$. Let $X = \{0\}, Y = \{1\}, Z = \{0, 1\}$. Then $\delta(X, Y) = 1$, $\delta(X, Z) = \delta(Y, Z) = \overline{\alpha}$.*

The proof of Lemma 9 is an immediate calculation.

Theorem 10. *$\delta_\alpha = \alpha\tilde{m} + \overline{\alpha}\tilde{M}$ satisfies the triangle inequality if and only if $0 \leq \alpha \leq 1/2$.*

Proof. We first show the *only if* direction. By Lemma 9 the triangle inequality only holds for the example given there if $1 \leq 2\overline{\alpha}$, i.e., $\alpha \leq 1/2$.

Now let us show the *if* direction. If $\alpha \leq 1/2$ then $\alpha \leq \overline{\alpha}$, so $\delta_\alpha = \alpha(\tilde{m} + \tilde{M}) + (\overline{\alpha} - \alpha)\tilde{M}$ is a nontrivial nonnegative linear combination. Since $(\tilde{m} + \tilde{M})(A, B) = |A \setminus B| + |B \setminus A|$ (Lemma 7) and $\tilde{M}(A, B) = \max\{|A \setminus B|, |B \setminus A|\}$ (Lemma 8) are both metrics, the result follows from Theorem 4. □

Lemma 11. *Suppose d is a metric on a collection of nonempty sets $\mathcal{X}$, with $d(X, Y) \leq 2$ for all $X, Y \in \mathcal{X}$. Let $\hat{\mathcal{X}} = \mathcal{X} \cup \{\emptyset\}$ and define $\hat{d} : \hat{\mathcal{X}} \times \hat{\mathcal{X}} \to \mathbb{R}$ by stipulating that for $X, Y \in \mathcal{X}$,*

$$\hat{d}(X, Y) = d(X, Y); \quad d(X, \emptyset) = 1 = d(\emptyset, X); \quad d(\emptyset, \emptyset) = 0.$$

Then $\hat{d}$ is a metric on $\hat{\mathcal{X}}$.

Theorem 12. *Let $f(A,B)$ be a metric such that*

$$|B \setminus A| \leq f(A,B)$$

for all A, B. Then the function d given by

$$d(A,B) = \begin{cases} \frac{f(A,B)}{|A\cap B|+f(A,B)}, & \text{if } |A\cap B| + f(A,B) > 0, \\ 0, & \text{otherwise}, \end{cases}$$

is a metric.

Proof. By Lemma 5 (with $a_{x,y} = |X \cap Y|$) we only need to verify that for all sets A, B, C,

$$|A\cap C| + f(A,B) \geq |B \cap C|.$$

And indeed, since tautologically $B \cap C \subseteq (B \setminus A) \cup (A \cap C)$, by the union bound we have $|B\cap C| - |A\cap C| \leq |B \setminus A| \leq f(A,B)$. □

Theorem 13. *Let $f(A,B) = m\min\{|A\setminus B|, |B\setminus A|\} + M\max\{|A\setminus B|, |B\setminus A|\}$ with $0 < m \leq M$ and $1 \leq M$. Then the function d given by*

$$d(A,B) = \begin{cases} \frac{f(A,B)}{|A\cap B|+f(A,B)}, & \text{if } A\cup B \neq \emptyset, \\ 0, & \text{otherwise}, \end{cases}$$

is a metric.

Proof. We have $f(A,B) = (m+M)\delta_\alpha(A,B)$ where $\alpha = \frac{m}{m+M}$. Since $m \leq M$, $\alpha \leq 1/2$, so f satisfies the triangle inequality by Theorem 10. Since $m > 0$, in fact f is a metric. Using $M \geq 1$,

$$f(A,B) \geq M\max\{|A\setminus B|, |B\setminus A|\} \geq M|B\setminus A| \geq |B\setminus A|,$$

so that by Theorem 12, d is a metric. □

0.1.3 Tversky indices

Definition 14 ([18])**.** For sets X and Y the Tversky index with parameters $\alpha, \beta \geq 0$ is a number between 0 and 1 given by

$$S(X,Y) = \frac{|X\cap Y|}{|X\cap Y| + \alpha|X\setminus Y| + \beta|Y\setminus X|}.$$

We also define the corresponding Tversky dissimilarity $d^T_{\alpha,\beta}$ by

$$d^T_{\alpha,\beta}(X,Y) = \begin{cases} 1 - S(X,Y) & \text{if } X\cup Y \neq \emptyset; \\ 0 & \text{if } X = Y = \emptyset. \end{cases}$$

Definition 15. The Szymkiewicz–-Simpson coefficient is defined by

$$\text{overlap}(X,Y) = \frac{|X\cap Y|}{\min(|X|,|Y|)}$$

We may note that $\text{overlap}(X,Y) = 1$ whenever $X \subseteq Y$ or $Y \subseteq X$, so that $1 - \text{overlap}$ is not a metric.

Definition 16. The Sørensen–Dice coefficient is defined by

$$\frac{2|X \cap Y|}{|X| + |Y|}.$$

Definition 17 ([7, Section 2])**.** Let $\mathcal{X}$ be a collection of finite sets. We define $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as follows. The symmetric Tversky ratio model is defined by

$$\mathbf{strm}(X, Y) = \frac{|X \cap Y| + \text{bias}}{|X \cap Y| + \text{bias} + \beta\left(\alpha\tilde{m} + (1-\alpha)\tilde{M}\right)}$$

The unbiased symmetric TRM (**ustrm**) is the case where bias $= 0$, which is the case we shall assume we are in for the rest of this paper. The Tversky semimetric $D_{\alpha,\beta}$ is defined by $D_{\alpha,\beta}(X, Y) = 1 - \mathbf{ustrm}(X, Y)$, or more precisely

$$D_{\alpha,\beta}(X, Y) = \begin{cases} \beta\frac{\alpha\tilde{m}+(1-\alpha)\tilde{M}}{|X\cap Y|+\beta(\alpha\tilde{m}+(1-\alpha)\tilde{M})}, & \text{if } X \cup Y \neq \emptyset; \\ 0 & \text{if } X = Y = \emptyset. \end{cases}$$

Note that for $\alpha = 1/2$, $\beta = 1$, the STRM is equivalent to the Sørensen–Dice coefficient. Similarly, for $\alpha = 1/2$, $\beta = 2$, it is equivalent to Jaccard's coefficient.

0.2 Tversky metrics

Theorem 18. *The function $D_{\alpha,\beta}$ is a metric only if $\beta \geq 1/(1-\alpha)$.*

Proof. Recall that with $D = D_{\alpha,\beta}$,

$$D(X, Y) = \frac{\beta\delta}{|X \cap Y| + \beta\delta}.$$

By Lemma 9, for the example given there we have

$$\begin{aligned} D(X, Y) &= \frac{\beta \cdot 1}{0 + \beta \cdot 1} = 1, \\ D(X, Z) = D(Y, Z) &= \frac{\beta \cdot \overline{\alpha}}{1 + \beta \cdot \overline{\alpha}}. \end{aligned}$$

The triangle inequality is then equivalent to:

$$1 \leq 2\frac{\beta\overline{\alpha}}{1 + \beta\overline{\alpha}} \iff \beta\overline{\alpha} \geq 1 \iff \beta \geq 1/(1-\alpha). \qquad \square$$

In Theorem 19 we use the interval notation on $\mathbb{N}$, given by $[a, a] = \{a\}$ and $[a, b] = [a, b-1] \cup \{b\}$.

Theorem 19. *The function $D_{\alpha,\beta}$ is a metric on all finite power sets only if $\alpha \leq 1/2$.*

Proof. Suppose $\alpha > 1/2$. Then $2\overline{\alpha} < 1$. Let n be an integer with $n > \frac{\beta\overline{\alpha}}{1-2\overline{\alpha}}$. Let $X_n = [0, n]$, and $Y_n = [1, n+1]$, and $Z_n = [1, n]$. The triangle inequality says

$$\begin{aligned} \beta\frac{1}{n + \beta \cdot 1} = D(X_n, Y_n) &\leq D(X_n, Z_n) + D(Z_n, Y_n) = 2\beta\frac{\overline{\alpha}}{n + \beta\overline{\alpha}} \\ n + \beta\overline{\alpha} &\leq 2\overline{\alpha}(n + \beta) \\ n(1 - 2\overline{\alpha}) &\leq \beta\overline{\alpha} \end{aligned}$$

Then the triangle inequality does not hold, so $D_{\alpha,\beta}$ is not a metric on the power set of $[0, n+1]$. □

Theorem 20. *Let $0 \leq \alpha \leq 1$ and $\beta > 0$. Then $D_{\alpha,\beta}$ is a metric if and only if $0 \leq \alpha \leq 1/2$ and $\beta \geq 1/(1-\alpha)$.*

Proof. Theorem 18 and Theorem 19 give the necessary condition. Since

$$D_{\alpha,\beta} = \begin{cases} \beta \frac{\alpha\tilde{m} + (1-\alpha)\tilde{M}}{|X \cap Y| + \beta(\alpha\tilde{m} + (1-\alpha)\tilde{M})}, & \text{if } X \cup Y \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{m}$ is the minimum of the set differences and $\tilde{M}$ is the maximum, we can let $f(A,B) = \beta(\alpha\tilde{m}(A,B) + (1-\alpha)\tilde{M}(A,B))$. Then with the constants $m = \beta\alpha$ and $M = \beta\overline{\alpha}$, we can apply Theorem 13. □

We have formally proved Theorem 20 in the Lean theorem prover. The Github repository can be found at [8].

0.2.1 A converse to Gragera and Suppakitpaisarn

Theorem 21 (Gragera and Suppakitpaisarn [5, 6])**.** *The optimal constant ρ such that $d^T_{\alpha,\beta}(X,Y) \leq \rho(d^T_{\alpha,\beta}(X,Y) + d^T_{\alpha,\beta}(Y,Z))$ for all X, Y, Z is*

$$\frac{1}{2}\left(1 + \sqrt{\frac{1}{\alpha\beta}}\right).$$

Corollary 22. *$d^T_{\alpha,\beta}$ is a metric only if $\alpha = \beta \geq 1$.*

Proof. Clearly, $\alpha = \beta$ is necessary to ensure $d^T_{\alpha,\beta}(X,Y) = d^T_{\alpha,\beta}(Y,X)$. Moreover $\rho \leq 1$ is necessary, so Theorem 21 gives $\alpha\beta \geq 1$. □

Theorem 23 gives the converse to the Gragera and Suppakitpaisarn inspired Corollary 22:

Theorem 23. *The Tversky dissimilarity $d^T_{\alpha,\beta}$ is a metric iff $\alpha = \beta \geq 1$.*

Proof. Suppose the Tversky dissimilarity $d^T_{\alpha,\beta}$ is a semimetric. Let X, Y be sets with $|X \cap Y| = |X \setminus Y| = 1$ and $|Y \setminus X| = 0$. Then

$$1 - \frac{1}{1+\beta} = d^T_{\alpha,\beta}(Y,X) = d^T_{\alpha,\beta}(X,Y) = 1 - \frac{1}{1+\alpha},$$

hence $\alpha = \beta$. Let $\gamma = \alpha = \beta$.

Now, $d^T_{\gamma,\gamma} = D_{\alpha_0,\beta_0}$ where $\alpha_0 = 1/2$ and $\beta_0 = 2\gamma$. Indeed, with $\tilde{m} = \min\{|X \setminus Y|, |Y \setminus X|\}$ and $\tilde{M} = \max\{|X \setminus Y|, |Y \setminus X|\}$, since

$$D_{\alpha_0,\beta_0} = \beta_0 \frac{\alpha_0\tilde{m} + (1-\alpha_0)\tilde{M}}{|X \cap Y| + \beta_0\left[\alpha_0\tilde{m} + (1-\alpha_0)\tilde{M}\right]},$$

$$D_{\frac{1}{2},2\gamma} = 2\gamma \frac{\frac{1}{2}\tilde{m} + (1-\frac{1}{2})\tilde{M}}{|X \cap Y| + 2\gamma\left[\frac{1}{2}\tilde{m} + (1-\frac{1}{2})\tilde{M}\right]}$$

$$= \gamma \frac{|X \setminus Y| + |Y \setminus X|}{|X \cap Y| + \gamma\,[|X \setminus Y| + |Y \setminus X|]} = 1 - \frac{|X \cap Y|}{|X \cap Y| + \gamma |X \setminus Y| + \gamma |Y \setminus X|} = d^T_{\gamma,\gamma}.$$

By Theorem 20, $d^T_{\gamma,\gamma}$ is a metric if and only if $\beta_0 \geq 1/(1-\alpha_0)$. This is equivalent to $2\gamma \geq 2$, i.e., $\gamma \geq 1$. □

The truth or falsity of Theorem 23 does not arise in Gragera and Suppakitpaisarn's work, as they require $\alpha, \beta \leq 1$ in their definition of Tversky index. We note that Tversky [18] only required $\alpha, \beta \geq 0$.

0.3 Lebesgue-style metrics

Incidentally, the names of J_1 and J_∞ come from the observation that they are special cases of J_p given by

$$J_p(A,B) = \left(2 \cdot \frac{|B \setminus A|^p + |A \setminus B|^p}{|A|^p + |B|^p + |B \setminus A|^p + |A \setminus B|^p}\right)^{1/p} = \begin{cases} J_1(A,B) & p = 1 \\ J_\infty(A,B) & p \to \infty \end{cases}$$

which was suggested in [9] as another possible means of interpolating between J_1 and J_∞. We still conjecture that J_2 is a metric, but shall not attempt to prove it here. However:

Theorem 24. *J_3 is not a metric.*

Because of Theorem 24, we searched for a better version of J_p, and found $\mathcal{V}_p$:

Definition 25. For each $1 \leq p \leq \infty$, let[1]

$$\begin{aligned} \Delta_p(A,B) &= (|B \setminus A|^p + |A \setminus B|^p)^{1/p}, \text{ and} \\ \mathcal{V}_p(A,B) &= \frac{\Delta_p(A,B)}{|A \cap B| + \Delta_p(A,B)}. \end{aligned}$$

We have $\mathcal{V}_1 = J_1$ and $\mathcal{V}_\infty := \lim_{p\to\infty} \mathcal{V}_p = J_\infty$.

In a way what is going on here is that we consider L^p spaces instead of

$$\frac{1}{p} L^1 + \left(1 - \frac{1}{p}\right) L^\infty$$

spaces.

Theorem 26. *For each $1 \leq p \leq \infty$, Δ_p is a metric.*

Theorem 27. *For each $1 \leq p \leq \infty$, $\mathcal{V}_p$ is a metric.*

Proof. By Theorem 26 and Theorem 12, we only have to check $|B \setminus A| \leq \Delta_p(A,B)$, which is immediate for $1 \leq p \leq \infty$. □

Of special interest may be $\mathcal{V}_2$ as a canonical interpolant between $\mathcal{V}_1$, the Jaccard distance, and $\mathcal{V}_\infty = J_\infty$, the analogue of the NID. If $|B \setminus A| = 3$, $|A \setminus B| = 4$, and $|A \cap B| = 5$, then

$$\begin{aligned} \mathcal{V}_1(A,B) &= 7/12, \\ \mathcal{V}_2(A,B) &= 1/2, \\ \mathcal{V}_\infty(A,B) &= 4/9. \end{aligned}$$

Note that if $A \subseteq B$ then $\mathcal{V}_p(A,B) = \mathcal{V}_1(A,B)$ for all p.

[1] Here, $\mathcal{V}$ can stand for Paul M. B. Vitányi, who introduced the author to the normalized information distance at a Dagstuhl workshop in 2006.

0.4 Conclusion and applications

Many researchers have considered metrics based on sums or maxima, but we have shown that these need not be considered in "isolation" in the sense that they form the endpoints of a family of metrics.

As an example, the mutations of spike glycoproteins of coronaviruses are of interest in connection with diseases such as CoViD-19. We calculated several distance measures between peptide sequences for such proteins. The distance

$$Z_{2,\alpha}(x_0, x_1) = \alpha \min(|A_1|, |A_2|) + \overline{\alpha} \max(|A_1|, |A_2|)$$

where A_i is the set of subwords of length 2 in x_i but not in x_{1-i}, counts how many subwords of length 2 appear in one sequence and not the other.
We used the Ward linkage criterion for producing Newick trees using the `hclust` package for the Go programming language. The calculated phylogenetic trees were based on the metric $Z_{2,\alpha}$.

We found one tree isomorphism class each for $0 \leq \alpha \leq 0.21$, $0.22 \leq \alpha \leq 0.36$, and $0.37 \leq \alpha \leq 0.5$, respectively. We see that the various intervals for α can correspond to "better" or "worse" agreement with other distance measures. Thus, we propose that rather than focusing on $\alpha = 0$ and $\alpha = 1/2$ exclusively, future work may consider the whole interval $[0, 1/2]$.

Bibliography

[1] R. Cilibrasi and P. M. B. Vitanyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.

[2] Rudi L. Cilibrasi and Paul M. B. Vitanyi. The Google similarity distance. *IEEE Trans. on Knowl. and Data Eng.*, 19(3):370–383, March 2007.

[3] Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Springer, Berlin, fourth edition, 2016.

[4] Ged Ridgway. Mutual information — Wikipedia, the Free Encyclopedia, Revision as of 14:55, 22 january 2010, 2010. [Online; accessed 14-May-2020].

[5] Alonso Gragera and Vorapong Suppakitpaisarn. Semimetric properties of Sørensen-Dice and Tversky indexes. In *WALCOM: algorithms and computation*, volume 9627 of *Lecture Notes in Comput. Sci.*, pages 339–350. Springer, [Cham], 2016.

[6] Alonso Gragera and Vorapong Suppakitpaisarn. Relaxed triangle inequality ratio of the Sørensen-Dice and Tversky indexes. *Theoret. Comput. Sci.*, 718:37–45, 2018.

[7] Sergio Jiménez, Claudia Jeanneth Becerra, and Alexander F. Gelbukh. SOFTCARDINALITY-CORE: improving text overlap with distributional measures for semantic textual similarity. In Mona T. Diab, Timothy Baldwin, and Marco Baroni, editors, *Proceedings of the Second Joint Conference on Lexical and Computational Semantics, *SEM 2013, June 13-14, 2013, Atlanta, Georgia, USA*, pages 194–201. Association for Computational Linguistics, 2013.

[8] Bjørn Kjos-Hanssen. Lean project: a 1-parameter family of metrics connecting jaccard distance to normalized information distance. https://github.com/bjoernkjoshanssen/jaccard, 2021.

[9] Bjørn Kjos-Hanssen, Saroj Niraula, and Soowhan Yoon. A parametrized family of Tversky metrics connecting the Jaccard distance to an analogue of the Normalized Information Distance. In Sergei Artemov and Anil Nerode, editors, *Logical Foundations of Computer Science*, pages 112–124, Cham, 2022. Springer International Publishing.

[10] A Kraskov, H Stögbauer, R. G Andrzejak, and P Grassberger. Hierarchical clustering using mutual information. *Europhysics Letters (EPL)*, 70(2):278–284, apr 2005.

[11] Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, and Peter Grassberger. Hierarchical clustering based on mutual information. *ArXiv*, q-bio.QM/0311039, 2003.

[12] Abraham Lempel and Jacob Ziv. On the complexity of finite sequences. *IEEE Trans. Inform. Theory*, IT-22(1):75–81, 1976.

[13] Ming Li, Jonathan H. Badger, Xin Chen, Sam Kwong, Paul E. Kearney, and Haoyong Zhang. An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, 17 2:149–54, 2001.

[14] Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul M. B. Vitányi. The similarity metric. *IEEE Trans. Inform. Theory*, 50(12):3250–3264, 2004.

[15] Edward Raff and Charles K. Nicholas. An Alternative to NCD for Large Sequences, Lempel–Ziv Jaccard Distance. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[16] C. Rajski. Entropy and metric spaces. In *Information theory (Symposium, London, 1960)*, pages 41–45. Butterworths, Washington, D.C., 1961.

[17] Suvrit Sra. Is the Jaccard distance a distance? MathOverflow. URL:https://mathoverflow.net/q/210750 (version: 2015-07-03).

[18] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327—352, 1977.

[19] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Trans. Inform. Theory*, IT-23(3):337–343, 1977.

[20] Jacob Ziv and Abraham Lempel. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inform. Theory*, 24(5):530–536, 1978.