A 1-parameter family of metrics
connecting Jaccard distance to normalized information distance

Bjørn Kjos-Hanssen

Computability Theory and Applications
Online Seminar

November 17, 2020

Joint work with Saroj Niraula, Soowhan Yoon, Sabrina Hardisty, Guanhong Li, and Jacqueline Millard.
This work was partially supported by grants from the Simons Foundation (#704836 to Bjørn Kjos-Hanssen) and Decision Research Corporation (University of Hawai'i Foundation Account #129-4770-4).

Abstract

Jiménez, Becerra, and Gelbukh (2013) defined a family of symmetric Tversky ratio models S parametrized by $0 \leq \alpha \leq 1$ and $\beta > 0$. Letting $D = 1 - S$ we have a semimetric which we show is a metric if and only if $0 \leq \alpha \leq \frac{1}{2}$ and $\beta \geq 1/(1-\alpha)$. For $\beta = 1/(1-\alpha)$, the two endpoints $\alpha = 0, \frac{1}{2}$ correspond to the normalized information distance and Jaccard distance, respectively.

Distance metrics are used in a wide variety of scientific contexts. In bioinformatics, M. Li, Badger, Chen, Kwong, and Kearney [LBC+01] introduced an information-based sequence distance. In an information-theoretical setting, M. Li, Chen, X. Li, Ma and Vitányi [LCL+04] rejected the distance of [LBC+01] in favor of a *normalized information distance* (NID):

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

where $K(x \mid y)$ is the prefix Kolmogorov complexity of x given y.

The fact that the NID is in a sense a normalized metric is proved in [LCL+04]. Then in 2017, while studying malware detection, Raff and Nicholas [RN17] suggested Lempel–Ziv Jaccard distance (LZJD) as a practical alternative to NID.
We show that the NID and Jaccard distances constitute the endpoints of a parametrized family of metrics.

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

$$\frac{|X \setminus Y| + |Y \setminus X|}{|X \cup Y|} = 1 - \frac{|X \cap Y|}{|X \cup Y|}, \quad \text{and} \tag{1}$$

$$\frac{\max\{|X \setminus Y|, |Y \setminus X|\}}{\max\{|X|, |Y|\}}, \tag{2}$$

Kraskov et al. [KSAG03, KSAG05] study 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 [Raj61].

STRM (Symmetric Tversky Ratio Models) are variants of the Tversky index proposed in [JBG13].

Definition

A semimetric on $\mathcal{X}$ is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies the first three axioms of a metric space, but not necessarily the triangle inequality: $d(x, y) \geq 0$, $d(x, y) = 0$ if and only if $x = y$, and $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$.

Definition

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

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

Lemma

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 $\hat{\mathcal{X}}$.

Definition

The Szymkiewicz—Simpson coefficient is defined by

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

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

Definition

The Sørensen–Dice coefficient is defined by

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

Definition ([JBG13, Section 2])

Let $\mathcal{X}$ be a collection of finite sets. We define $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as follows. For sets $X, Y \in \mathcal{X}$ we define
$m(X, Y) = \min\{|X \setminus Y|, |Y \setminus X|\}$ and
$M(X, Y) = \max\{|X \setminus Y|, |Y \setminus X|\}$. The symmetric TRM is defined by

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

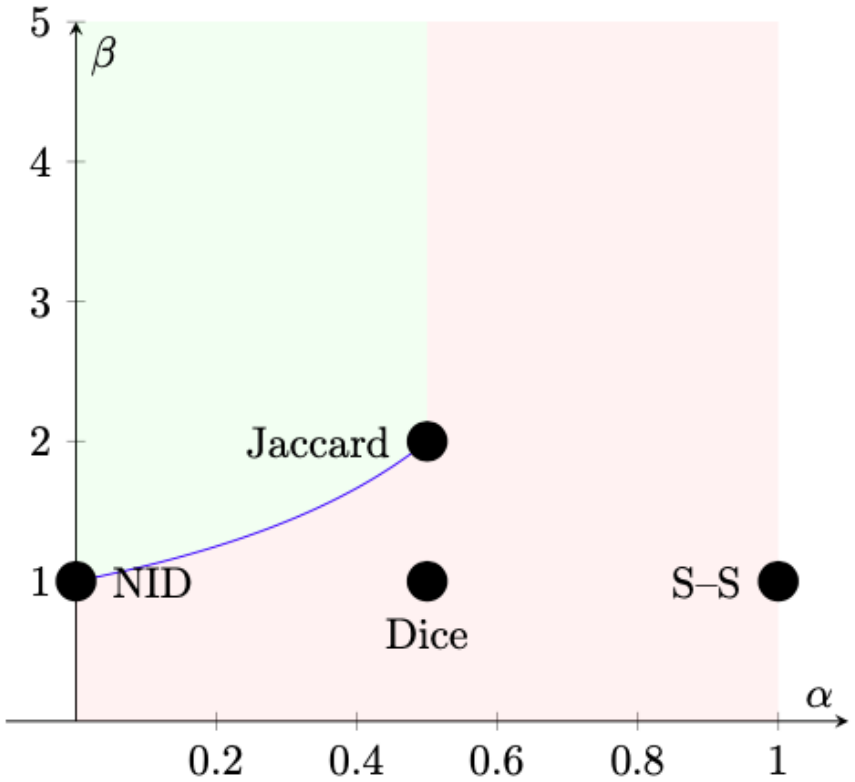
The unbiased symmetric TRM is the case where $\text{bias} = 0$, which is the case we shall assume. The Tversky semimetric $D'_{\alpha,\beta}$ is defined by $D'_{\alpha,\beta}(X, Y) = 1 - S(X, Y)$, or more precisely

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

Theorem

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* $\beta \geq 1/(1-\alpha)$.

A 1-parameter family of metrics



Lemma

Let $u, v, w, \epsilon > 0$. Then

$$\frac{1}{u} \leq \frac{1}{v} + \frac{1}{w} \rightarrow \frac{1}{u+\epsilon} \leq \frac{1}{v+\epsilon} + \frac{1}{w+\epsilon}.$$

Lemma

Suppose $a(x,y) = a_{xy}$ *and* $b(x,y) = b_{xy}$ *are functions. Suppose the function* d *given by* $d(x,y) = a_{xy}/b_{xy}$ *is a metric, and* $\epsilon > 0$ *is a real number. Let* $\hat{d}(x,y) = \frac{a_{xy}}{b_{xy} + \varepsilon a_{xy}}$. *Then* $\hat{d}$ *is also a metric.*

Theorem

For each α, the set of β for which $D'_{\alpha,\beta}$ is a metric is upward closed.

Some convenient notation: $\overline{\alpha} = 1 - \alpha$; $x_{\cap}y = |X \cap Y|$, $x = |X|$; $x_y = |X \setminus Y|$, $x_{zy} = |X \setminus (Z \cup Y)| = |(X \setminus Z) \setminus Y|$.

Theorem

$\delta := \alpha m + \overline{\alpha} M$ satisfies the triangle inequality if and only if $\alpha \leq 1/2$.

The result was conjectured/discovered using `Python` and the proof was verified in `Lean`.
It uses the identity $x_y + y_z + z_x = x_z + z_y + y_x$ which holds generally since both sides counts the elements that belong to exactly one of X, Y, Z once each, and counts the elements that belong to exactly two of X, Y, Z once each.

Theorem

The function $D'_{\alpha,\beta}$ is a metric only if $\beta \geq 1/(1-\alpha)$.

Theorem

The function $D'_{\alpha,\beta}$ is a metric on all finite power sets only if $\alpha \leq 1/2$.

Can there be α and β with $\beta \geq 1/(1-\alpha)$ and $0 \leq \alpha \leq 1/2$ such that in terms of

$$\delta = (1-\alpha)\max\{K(x \mid y), K(y \mid x)\} + \alpha \min\{K(x \mid y), K(y \mid x)\},$$

and

$$I(X,Y) = K(X) + K(Y) - K(X,Y),$$

$$D'_{\alpha,\beta} = \frac{\beta\delta}{I(X,Y) + \beta\delta}$$

is lower or upper semicomputable?

Terwijn et al. [TTV11] showed the answer is No in the case $\alpha = 0$ and $\beta = 1$, the NID:

$$D'_{0,1} = \frac{\max\{K(x \mid y), K(y \mid x)}{K(X) + K(Y) - K(X, Y) + \max\{K(x \mid y), K(y \mid x)}$$

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

Entropy version of main result

McGill [McG54] and Hu Kuo Ting [Tin62] established a connection between Venn diagram, signed measures, and entropies.

Lemma

The cardinalities of the 7 minimal regions of the Venn diagram of X, Y, Z are definable from $+$ and $-$ given the cardinalities of $X \cup Y \cup Z$, $X \cup Y$, $X \cup Z$, $Y \cup Z$, X, Y, Z.

Proof.

We have $|X \setminus Y| = |X \cup Y| - |Y|$,
$|X \cap Y| = |X \cup Y| - |X \setminus Y| - |Y \setminus X|$, and
$|X \cup Y \cup Z| = |X| + |Y| + |Z| - |X \cap Y| - |X \cap Z| - |Y \cap Z| + |X \cap Y \cap Z|$,
so $|X \cap Y \cap Z|$ is definable. □

Entropy version of main result

Using the Lemma we can define conditional mutual information $I(X;Y;Z)$, $I(X;Y \mid Z)$ etc. from joint entropies $H(X,Y,Z)$, $H(X,Y)$, etc. The only region of the Venn diagram that may be negative is $I(X;Y;Z)$, but $|X \cap Y \cap Z| \geq 0$ was never used in the proof of the main theorem above.

By definition $I(x;y;z) = I(x;y) - I(x;y \mid z)$ and
$I(x;y) = H(x,y) - H(y \mid x) - H(x \mid y)$ and
$I(x;y \mid z) = H(x,z) + H(y,z) - H(x,y,z) - H(z)$
The semicolon serves to separate random variables whereas a comma puts them together. Thus $I(X;Y,Z)$ is the mutual information between X and the vector (Y,Z).

Definition

For a finite set Ω, let $\mathbb{R}^\Omega$ be the set of real-valued functions on Ω and let $P(\Omega)$ be the set of all probability distributions on Ω. For each $\mathbb{P} \in P(\Omega)$, $RV(\mathbb{P}) = (\mathbb{R}^\Omega, \mathbb{P})$ is a space of random variables.

We call $RV(\mathbb{P})$ a "space" since random variables may be added, multiplied, etc., although that fact is not used in the following theorem.

Theorem

Let

$$\delta(X,Y) = \alpha \min\{H(X \mid Y), H(Y \mid X)\}$$
$$+(1-\alpha)\max\{H(X \mid Y), H(Y \mid X)\}.$$

Let

$$D'_{\alpha,\beta}(X,Y) = \begin{cases} \beta \frac{\delta(X,Y)}{I(X;Y)+\beta(\delta(X,Y))}, & H(X,Y) > 0, \\ 0, & H(X,Y) = 0. \end{cases}$$

The following are equivalent.

- $D'_{\alpha,\beta}$ *is a metric on all spaces* $RV(\Omega)$;
- $0 \leq \alpha \leq 1/2$ *and* $\beta \geq 1/(1-\alpha)$.

Theorem

Let X_1, X_2, X_3 be random variables. The following quantities are nonnegative: $H(X_i), H(X_i \cup X_j), H(X_1 \cup X_2 \cup X_3)$, $H(X_i) + H(X_j) - H(X_i \cup X_j)$, ... but not necessarily $I(X_1; X_2; X_3) = ...$

See [CT06, Chapter 2, Theorem 2.6.3, Exercise 2.29]. Jensen's Inequality is used to show $I(X; Y) \geq 0$.

Negative mutual information

We have negative "mutual information" $I(X;Y;Z)$. for $X = 1001010100010110, Y = 0110110110100011, Z = 1111100010110101$.

However, using Watanabe's [Wat60] total correlation instead, $\sum H(X_i) - H(X_1, \ldots, X_k)$, the answer is positive.
Romaschenko and Zimand [RZ19] state that "mutual information is only defined for two strings".

Note that for entropy purposes, random variables are in 1–1 correspondence with partitions of Ω.
When embedding semilattices into the Turing degrees we use representations of lattices by equivalence relations.
This means that "high entropy" correspond to "high Turing degree".

Note that for entropy purposes, random variables are in 1–1 correspondence with partitions of Ω.
When embedding semilattices into the Turing degrees we use representations of lattices by equivalence relations.
This means that "high entropy" correspond to "high Turing degree".

For the constructed functions $g^{[k]}$ in initial segment constructions, it is clear that $C(g^{[i]} \upharpoonright n) \leq C(g^{[j]} \upharpoonright n)$ when $i \leq j$ in the lattice, simply because the use is identity-bounded.
This suggests that the Turing degrees of unsolvability are ordered the "correct" way, i.e., we should not put $\mathbf{0}$ on top because it is the "most solvable".

Representations of lattices originate by equivalence relations are of interest when studying the lattice of all subalgebras or quotient algebras of a structure.
A quotient of a free structure can be a complicated structure, and a quotient of a complicated structure can be a singleton. The identity relation and the "all" relation are both "simple". The point is that the identity relation has higher entropy (no matter the distribution).


Thomas M. Cover and Joy A. Thomas.
Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).
Wiley-Interscience, USA, 2006.


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.


Alexander Kraskov, Harald Stögbauer, Ralph G. Andrzejak, and Peter Grassberger.

Hierarchical clustering based on mutual information.
ArXiv, q-bio.QM/0311039, 2003.


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


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.


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.


William J. McGill.
Multivariate information transmission.

Psychometrika, 19(2):97–116, 1954.


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


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.


Andrei Romashchenko and Marius Zimand.
An operational characterization of mutual information in algorithmic information theory.
J. ACM, 66(5), September 2019.


Hu Kuo Ting.

On the amount of information.
Theory of Probability and Its Applications, 7:439–447, 1962.


Sebastiaan A. Terwijn, Leen Torenvliet, and Paul M. B. Vitányi.
Nonapproximability of the normalized information distance.
J. Comput. System Sci., 77(4):738–742, 2011.


S. Watanabe.
Information theoretical analysis of multivariate correlation.
IBM Journal of Research and Development, 4(1):66–82, 1960.