Theoretical Computer Science 410 (2009) 2365–2376

Contents lists available at ScienceDirect

Theoretical Computer Science

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

On the similarity metric and the distance metric[☆]

Shihyen Chen, Bin Ma, Kaizhong Zhang[*]

Department of Computer Science, The University of Western Ontario, London, Ontario, Canada, N6A 5B7

ARTICLE INFO

Keywords:
Similarity metric
Distance metric
Normalized similarity metric
Normalized distance metric

ABSTRACT

Similarity and dissimilarity measures are widely used in many research areas and applications. When a dissimilarity measure is used, it is normally required to be a distance metric. However, when a similarity measure is used, there is no formal requirement. In this article, we have three contributions. First, we give a formal definition of similarity metric. Second, we show the relationship between similarity metric and distance metric. Third, we present general solutions to normalize a given similarity metric or distance metric.

1. Introduction

Distance and similarity measures are widely used in bioinformatics research and other fields. Here, we give a few examples.

Distance: Sequence edit distance and tree edit distance are used in many areas [14,28]. Moreover, the widespread use of distance is exemplified in the following contexts: constructing phylogenetic trees [19,21,23], improving database search [18], describing the relationship between words [3], comparing graphs or attributed trees [2,24], comparing information contents [8], and evaluating the importance of attributes in data mining [10,17,25].

Similarity: Protein sequence similarity based on BLOSUM matrices is used for protein sequence comparison [20]. Similarity metrics are used in data mining for evaluating the importance of attributes [5–7,9,12,16].

Distance metric is a well defined concept. In contrast, although similarity measures are widely used and their properties are studied and discussed [22,26], it seems that there is no formal definition for the concept. In this article, we give a formal definition of similarity metric. We then show the relationship between similarity metric and distance metric. Furthermore, we consider the problem of normalized similarity metric and normalized distance metric. Although there are studies on normalizing specific similarity and distance metrics [2,5,7,8,10,12,17,24], there is no general solution. We present general solutions to normalize a given similarity metric or distance metric. Finally, we illustrate with examples the generality of the presented solutions.

The rest of the paper is organized as follows. Section 2 reviews the definition of distance metric and introduces a formal definition of similarity metric while showing some useful properties. Section 3 concerns the relationship between similarity metric and distance metric. Section 4 concerns normalized similarity metric. Section 5 concerns normalized distance metric. Section 6 concerns the generality of the presented solutions. Section 7 presents concluding remarks.

[☆] Preliminary work can be found in [S. Chen, B. Ma, K. Zhang, The normalized similarity metric and its applications, in: Proceedings of 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007, 2007, pp. 172–180; B. Ma, K. Zhang, The similarity metric and the distance metric, in: Proceedings of the 6th Atlantic Symposium on Computational Biology and Genome Informatics, 2005, pp. 1239–1242]. This work was partially supported by NSERC grants.

[*] Corresponding author.
E-mail addresses: schen@csd.uwo.ca (S. Chen), bma@csd.uwo.ca (B. Ma), kzhang@csd.uwo.ca (K. Zhang).

0304-3975/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2009.02.023

2. Similarity metric and distance metric

Recall the formal definition of a distance metric as follows.

Definition 1 (*Distance Metric*). Given a set X, a real-valued function $d(x, y)$ on the Cartesian product $X \times X$ is a distance metric if for any $x, y, z \in X$, it satisfies the following conditions:

1. $d(x, y) \geq 0$ (non-negativity),
2. $d(x, y) = d(y, x)$ (symmetry),
3. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality),
4. $d(x, y) = 0$ if and only if $x = y$ (identity of indiscernibles).

To our knowledge, there is no formal metric definition for similarity. In the following, we present a formal definition for the similarity metric [4,11].

Definition 2 (*Similarity Metric*). Given a set X, a real-valued function $s(x, y)$ on the Cartesian product $X \times X$ is a similarity metric if, for any $x, y, z \in X$, it satisfies the following conditions:

1. $s(x, y) = s(y, x)$,
2. $s(x, x) \geq 0$,
3. $s(x, x) \geq s(x, y)$,
4. $s(x, y) + s(y, z) \leq s(x, z) + s(y, y)$,
5. $s(x, x) = s(y, y) = s(x, y)$ if and only if $x = y$.

Condition 1 states that $s(x, y)$ is symmetric. Condition 2 states that for any x the self similarity is nonnegative. Although it is not mandatory to set this lower bound at zero, this is a common and reasonable choice. Condition 3 states that for any x the self similarity is no less than the similarity between x and any y. Condition 4 states that the similarity between x and z through y is no greater than the direct similarity between x and z plus the self similarity of y. This property is the equivalent of the triangle inequality in distance metric. Condition 5 states that the statements $s(x, x) = s(y, y) = s(x, y)$ and $x = y$ are equivalent.

With the possible exceptions of condition 4 and 5, the remaining conditions clearly agree with the intuitive meaning of similarity. As to condition 4 and 5, although their relevance to similarity may not be intuitively clear, we explain in the following that they are indeed indispensable properties for similarity.

Consider condition 4. At first sight, this inequality might appear unnatural since, from the triangle inequality, one might expect it to be $s(x, y) + s(y, z) \leq s(x, z)$ without the $s(y, y)$ term. In a deeper analysis, as follows, we shall see why this $s(y, y)$ term should be included.

Intuitively, the notion of similarity serves as a means to quantify the common information shared by two objects. Two scenarios arise. In the first scenario, only non-negative values are used to quantify similarity. In the second scenario, real values are used to quantify similarity. In the current discussion, we borrow notations from set theory due to its convenience in conveying the intuition underlying similarity.

For non-negative quantification, the similarity between x and y may be expressed as $|x \cap y|$ which represents that which is commonly shared by both objects. Moreover, note that $|x \cap y| = |x \cap y \cap z| + |x \cap y \cap \bar{z}|$, where a notation $\bar{x}$ denotes the complement of x. In this scenario, we are concerned with the inequality:

$$|x \cap y| + |y \cap z| \leq |x \cap z| + |y| .$$

The validity of this inequality is justified as

$$|x \cap y| + |y \cap z| = |x \cap y \cap z| + |x \cap y \cap \bar{z}| + |x \cap y \cap z| + |\bar{x} \cap y \cap z| \leq |x \cap z| + |y| ,$$

due to the facts that $|x \cap y \cap z| \leq |x \cap z|$ and $|x \cap y \cap \bar{z}| + |x \cap y \cap z| + |\bar{x} \cap y \cap z| \leq |y|$. Without the presence of $|y|$, one cannot say that $|x \cap z|$ alone is enough to bound all the terms on the other side of the inequality. A simple example is when $|x \cap z| = \emptyset$ while $|x \cap y| \neq \emptyset$ and $|y \cap z| \neq \emptyset$.

For general quantification, the similarity between x and y may be expressed as

$$k \times |x \cap y| - k' \times (|x \cap \bar{y}| + |y \cap \bar{x}|)$$

where both common and non-common contributions are taken into account. In this scenario, we are concerned with the inequality:

$$k \times (|x \cap y| + |y \cap z|) - k' \times (|x \cap \bar{y}| + |y \cap \bar{x}| + |y \cap \bar{z}| + |z \cap \bar{y}|) \leq k \times (|x \cap z| + |y|) - k' \times (|x \cap \bar{z}| + |z \cap \bar{x}|).$$

From the results in the non-negative quantification, if we can show the validity of the following inequality then the validity of the above inequality follows:

$$|x \cap \bar{y}| + |y \cap \bar{x}| + |y \cap \bar{z}| + |z \cap \bar{y}| \geq |x \cap \bar{z}| + |z \cap \bar{x}| .$$

As shown in the following, this is indeed true:

$$\begin{aligned}|x\cap\bar{y}|+|y\cap\bar{x}|+|y\cap\bar{z}|+|z\cap\bar{y}| &\geq |x\cap\bar{y}\cap\bar{z}|+|\bar{x}\cap y\cap z|+|x\cap y\cap\bar{z}|+|\bar{x}\cap\bar{y}\cap z|\\ &= (|x\cap y\cap\bar{z}|+|x\cap\bar{y}\cap\bar{z}|)+(|\bar{x}\cap y\cap z|+|\bar{x}\cap\bar{y}\cap z|)\\ &= |x\cap\bar{z}|+|z\cap\bar{x}|\,.\end{aligned}$$

Now consider condition 5. The "if" part is clear. The "only-if" part, which states that if $s(x,x)=s(y,y)=s(x,y)$ then $x=y$, is justified by Lemma 1.

Lemma 1. *Let $s(x,y)$ be a real function satisfying similarity metric conditions 1, 2, 3 and 4. If $s(x,x)=s(y,y)=s(x,y)$ then for any z, $s(x,z)=s(y,z)$.*

Proof. From $s(x,y)+s(y,z)\leq s(x,z)+s(y,y)$, we have $s(y,z)\leq s(x,z)$. From $s(y,x)+s(x,z)\leq s(y,z)+s(x,x)$, we have $s(x,z)\leq s(y,z)$. This means that for any z, $s(x,z)=s(y,z)$. □

From the definitions, the negation of a distance metric is a similarity metric. Therefore the similarity metric is a more general notion. The next two lemmas consider the result of adding or multiplying two similarity metrics.

Lemma 2. *Let $s_1(x,y)\geq 0$ and $s_2(x,y)\geq 0$ be two similarity metrics, then $s_1(x,y)+s_2(x,y)$ is a similarity metric.*

Proof. Trivial. □

Lemma 3. *Let $s_1(x,y)\geq 0$ and $s_2(x,y)\geq 0$ be two similarity metrics, then $s_1(x,y)\times s_2(x,y)$ is a similarity metric.*

Proof. We only show the proof for condition 4 as the other conditions can be proved trivially.
Condition 4: Let $d_{xz}=\max\{s_2(x,y)+s_2(y,z)-s_2(y,y),0\}$, then $d_{xz}\leq s_2(x,z)$, $d_{xz}\leq s_2(y,y)$ and $s_2(x,y)+s_2(y,z)\leq s_2(y,y)+d_{xz}$. Without loss of generality, we assume that $s_1(x,y)\geq s_1(y,z)$. Then,

$$\begin{aligned}s_1(x,y)\times s_2(x,y)+s_1(y,z)\times s_2(y,z) &= (s_1(x,y)-s_1(y,z))\times s_2(x,y)+s_1(y,z)\times(s_2(x,y)+s_2(y,z))\\ &\leq (s_1(x,y)-s_1(y,z))\times s_2(y,y)+s_1(y,z)\times(s_2(y,y)+d_{xz})\\ &= s_1(x,y)\times(s_2(y,y)-d_{xz})+(s_1(x,y)+s_1(y,z))\times d_{xz}\\ &\leq s_1(y,y)\times(s_2(y,y)-d_{xz})+(s_1(y,y)+s_1(x,z))\times d_{xz}\\ &\leq s_1(y,y)\times s_2(y,y)+s_1(x,z)\times s_2(x,z).\quad\square\end{aligned}$$

Following the definitions of distance and similarity, the normalized metrics are defined as follows.

Definition 3 (*Normalized Distance Metric*). A distance metric $d(x,y)$ is a normalized distance metric if $d(x,y)\leq 1$.

Definition 4 (*Normalized Similarity Metric*). A similarity metric $s(x,y)$ is a normalized similarity metric if $|s(x,y)|\leq 1$.

Corollary 1. *If $s(x,y)$ is a normalized similarity metric and for any x, $s(x,x)=1$, then $\frac{1}{2}\times(1-s(x,y))$ is a normalized distance metric. If, in addition, $s(x,y)\geq 0$, then $1-s(x,y)$ is a normalized distance metric. If $d(x,y)$ is a normalized distance metric, then $1-d(x,y)$ is a normalized similarity metric.*

Proof. The statements follow directly from the basic definitions. □

Therefore, if $d_i(x,y)\geq 0$, $1\leq i\leq n$, are normalized distance metrics, then $\prod_i^n(1-d_i(x,y))$ is a normalized similarity metric and $1-\prod_i^n(1-d_i(x,y))$ is a normalized distance metric.

In the following, we discuss some properties of concave and convex functions that will be useful later.

Definition 5. A function f is concave over an interval $[a,b]$ if for every $x_1,x_2\in[a,b]$ and $0\leq\lambda\leq 1$,

$$\lambda\times f(x_1)+(1-\lambda)\times f(x_2)\leq f(\lambda\times x_1+(1-\lambda)\times x_2). \tag{1}$$

Definition 6. A function f is convex over an interval $[a,b]$ if for every $x_1,x_2\in[a,b]$ and $0\leq\lambda\leq 1$,

$$\lambda\times f(x_1)+(1-\lambda)\times f(x_2)\geq f(\lambda\times x_1+(1-\lambda)\times x_2).$$

Lemma 4. *If a function f is concave over interval $(-\infty,\infty)$, then for any $a,b\geq 0$ and $c\geq 0$,*

$$f(a)+f(a+b+c)\leq f(a+b)+f(a+c).$$

Proof. Let $a+b=\lambda\times a+(1-\lambda)\times(a+b+c)$ and $a+c=\lambda'\times a+(1-\lambda')\times(a+b+c)$. Consequently $\lambda=\frac{c}{b+c}$, $\lambda'=\frac{b}{b+c}$, and $\lambda+\lambda'=1$. From Eq. (1), we have

$$\begin{cases}\lambda\times f(a)+(1-\lambda)\times f(a+b+c)\leq f(\lambda\times a+(1-\lambda)\times(a+b+c))=f(a+b),\\ \lambda'\times f(a)+(1-\lambda')\times f(a+b+c)\leq f(\lambda'\times a+(1-\lambda')\times(a+b+c))=f(a+c).\end{cases}$$

Hence, $f(a)+f(a+b+c)\leq f(a+b)+f(a+c)$. □

Lemma 5. *If a function f is convex over interval $(-\infty, \infty)$, then for any $a, b \geq 0$ and $c \geq 0$,*

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

Proof. Symmetric to Lemma 4. □

Lemma 6. *Let f be a non-negative concave function over $[0, \infty)$. Then $\frac{x}{f(b+x)} \leq \frac{y}{f(b+y)}$ where $0 \leq x \leq y$ and $0 \leq b$.*

Proof. Let $0 \leq \lambda \leq 1$ such that $\lambda \times b + (1-\lambda) \times (b+y) = b+x$. Then $(1-\lambda) \times y = x$ and

$$\frac{f(b+x)}{x} \geq \frac{\lambda \times f(b) + (1-\lambda) \times f(b+y)}{x} = \frac{\lambda \times f(b)}{x} + \frac{f(b+y)}{y} \geq \frac{f(b+y)}{y}.$$

Hence $\frac{x}{f(b+x)} \leq \frac{y}{f(b+y)}$. □

The next lemma states the consequence of setting a similarity metric as an argument of a convex function.

Lemma 7. *Let $s(x, y)$ be a similarity metric, and f a convex function such that $f(0) \geq 0$, and $f(x) < f(y)$ if $x < y$. Then $f(s(x, y))$ is a similarity metric.*

Proof.
Condition 1, 2 and 3: Trivial.
Condition 4: Let $a = s(x, y) + s(y, z) - s(y, y)$, $b = s(y, y) - s(x, y)$ and $c = s(y, y) - s(y, z)$. Then it is straightforward to verify this condition with the help of Lemma 5.
Condition 5: If $x = y$ then clearly $f(s(x, x)) = f(s(y, y)) = f(s(x, y))$. Conversely, $f(s(x, x)) = f(s(y, y)) = f(s(x, y))$ implies $s(x, x) = s(y, y) = s(x, y)$ due to the condition that $f(x) < f(y)$ if $x < y$, hence $x = y$. □

Note. If the functional condition in Lemma 7 becomes "$f(x) \leq f(y)$ if $x < y$", then by partitioning the set into equivalence classes such that x and y are in the same class if and only if $f(s(x, x)) = f(s(y, y)) = f(s(x, y))$, $f(s(x, y))$ is still a similarity metric on the quotient set.

Corollary 2. *Given a similarity metric $s(x, y)$ on X, we define $s^+(x, y)$ as follows.*

$$s^+(x, y) = \begin{cases} s(x, y), & s(x, y) \geq 0, \\ 0, & s(x, y) < 0. \end{cases}$$

Then $s^+(x, y)$ is a similarity metric on X' where all $x \in X$ such that $s(x, x) = 0$ correspond to a single element in X'.

Proof. The result follows directly from the preceding note. □

3. Relationship between similarity metric and distance metric

We consider the relationship between the similarity metric and the distance metric. In particular, we establish transformations that transform a given similarity metric to a distance metric and vice versa.

We first consider transformations from similarity metric to distance metric. Given a similarity metric $s(x, y)$, we define two transformations, $F_p(s) = d_p$ and $F_m(s) = d_m$, as follows:

$$F_p(s(x, y)) = \frac{s(x, x) + s(y, y)}{2} - s(x, y),$$
$$F_m(s(x, y)) = \max\{s(x, x), s(y, y)\} - s(x, y).$$

In the following, we prove that these transformations produce distance metrics.

Lemma 8. *Let $s(x, y)$ be a similarity metric. Then,*

$$d_p(x, y) = \frac{s(x, x) + s(y, y)}{2} - s(x, y)$$

is a distance metric.

Proof.
Condition 1:

$$d_p(x, y) = \frac{s(x, x) + s(y, y)}{2} - s(x, y) = \frac{s(x, x) - s(x, y) + s(y, y) - s(x, y)}{2} \geq 0.$$

The inequality is due to similarity metric condition 3.
Condition 2: Trivial.

Condition 3:

$$\begin{aligned} d_p(x,z) &= \frac{s(x,x)+s(z,z)-2\times s(x,z)}{2} \\ &\le \frac{s(x,x)+s(z,z)+2\times s(y,y)-2\times s(x,y)-2\times s(y,z)}{2} \\ &= \frac{s(x,x)+s(y,y)-2\times s(x,y)}{2}+\frac{s(y,y)+s(z,z)-2\times s(y,z)}{2} \\ &= d_p(x,y)+d_p(y,z). \end{aligned}$$

Condition 4: If $x = y$ then clearly $d_p(x,y) = 0$. Conversely, $d_p(x,y) = 0$ means $s(x,x)+s(y,y)-2\times s(x,y) = 0$. Since $s(x,x) \ge s(x,y)$ and $s(y,y) \ge s(x,y)$, we must have $s(x,x) = s(x,y)$ and $s(y,y) = s(x,y)$ for $s(x,x)+s(y,y)-2\times s(x,y) = 0$ to hold, that is, $s(x,x) = s(y,y) = s(x,y)$. Hence, $x = y$. □

Lemma 9. *Let $s(x,y)$ be a similarity metric. Then,*

$$d_m(x,y) = \max\{s(x,x), s(y,y)\} - s(x,y)$$

is a distance metric.

Proof.
Condition 1 and 2: Trivial.
Condition 3:

$$\begin{aligned} d_m(x,z) &= \max\{s(x,x), s(z,z)\} - s(x,z) \\ &\le \max\{s(x,x), s(z,z)\} + s(y,y) - s(x,y) - s(y,z) \\ &\le \max\{s(x,x), s(y,y)\} - s(x,y) + \max\{s(y,y), s(z,z)\} - s(y,z) \\ &= d_m(x,y) + d_m(y,z). \end{aligned}$$

Condition 4: If $x = y$, then clearly $d_m(x,y) = 0$. Conversely, $d_m(x,y) = 0$ means $\max\{s(x,x), s(y,y)\} - s(x,y) = 0$. Since $s(x,x) \ge s(x,y)$ and $s(y,y) \ge s(x,y)$, this implies $s(x,x) = s(y,y) = s(x,y)$, hence $x = y$. □

Next, we consider transformations from distance metric to similarity metric. Given a distance metric $d(x,y)$ on X, we define, for any fixed $o \in X$, transformations $G_p^k(d) = s_p^k$ with $k \ge 1$, and $G_m^k(d) = s_m^k$ with $k > 0$, as follows:

$$\begin{aligned} G_p^k(d(x,y)) &= \frac{d(x,o)+d(y,o)}{k} - d(x,y), \\ G_m^k(d(x,y)) &= k \times \min\{d(x,o), d(y,o)\} - d(x,y). \end{aligned}$$

In the following, we prove that these transformations produce similarity metrics.

Lemma 10. *Let $d(x,y)$ be a distance metric on X. Then for $k \ge 1$, and any fixed $o \in X$,*

$$s_p^k(x,y) = \frac{d(x,o)+d(y,o)}{k} - d(x,y)$$

is a similarity metric.

Proof.
Condition 1, 2, 3 and 4: Trivial.
Condition 5: If $x = y$ then $s_p^k(x,x) = s_p^k(y,y) = s_p^k(x,y)$ holds trivially. Conversely, $s_p^k(x,x) = s_p^k(y,y) = s_p^k(x,y)$ implies $2 \times d(x,o) = 2 \times d(y,o) = d(x,o) + d(y,o) - k \times d(x,y)$. This means that $d(x,o) = d(y,o)$ and therefore $2 \times d(x,o) = 2 \times d(x,o) - k \times d(x,y)$. This yields $d(x,y) = 0$, hence $x = y$. □

Lemma 11. *Let $d(x,y)$ be a distance metric on X. Then for $k > 0$, and any fixed $o \in X$,*

$$s_m^k(x,y) = k \times \min\{d(x,o), d(y,o)\} - d(x,y)$$

is a similarity metric.

Proof.
Condition 1, 2 and 3: Trivial.
Condition 4:

$$\begin{aligned} s_m^k(x,y) + s_m^k(y,z) &= k \times \min\{d(x,o), d(y,o)\} - d(x,y) + k \times \min\{d(y,o), d(z,o)\} - d(y,z) \\ &\le k \times \min\{d(x,o), d(y,o)\} - d(x,z) + k \times \min\{d(y,o), d(z,o)\} - d(y,y) \\ &\le k \times \min\{d(x,o), d(z,o)\} - d(x,z) + k \times \min\{d(y,o), d(y,o)\} - d(y,y) \\ &= s_m^k(x,z) + s_m^k(y,y). \end{aligned}$$

Condition 5: If $x = y$ then $s_m^k(x,x) = s_m^k(y,y) = s_m^k(x,y)$ clearly holds. Conversely, $s_m^k(x,x) = s_m^k(y,y) = s_m^k(x,y)$ implies $k \times d(x,o) = k \times d(y,o) = k \times \min\{d(x,o), d(y,o)\} - d(x,y)$. This means $d(x,y) = 0$, hence $x = y$. □

Note. Given a distance metric d, we have $F_p(G_p^k(d)) = d$. Given a similarity metric s, in general $G_p^k(F_p(s)) \neq s$. Only when there exists a fixed $o \in X$, such that $(k-1) \times (s(x,x) + s(y,y)) = 2 \times (s(o,o) - s(x,o) - s(y,o))$, we have $G_p^k(F_p(s)) = s$.

The following lemma states a result involving transformation via exponential function.

Lemma 12. *If $d(x,y)$ is a distance metric, then $e^{-d(x,y)}$ is a normalized similarity metric and $1 - e^{-d(x,y)}$ is a normalized distance metric.*

Proof. From $(1 - e^{-d(x,y)}) \times (1 - e^{-d(y,z)}) \geq 0$, we have $e^{-d(x,y)} + e^{-d(y,z)} \leq e^{-(d(x,y)+d(y,z))} + 1$. Therefore $e^{-d(x,y)} + e^{-d(y,z)} \leq e^{-d(x,z)} + e^{-d(y,y)}$. Other properties are trivial. □

4. Normalized similarity metric

We first present several similarity metrics with a normalized appearance but which may not be strictly normalized according to Definition 4. Following these, we strengthen the functional condition so as to normalize these metrics.

Theorem 1. *Let $s(x,y)$ be a similarity metric, and f a concave function over $[0,\infty)$ satisfying $f(0) \geq 0$, $f(x) > 0$ if $x > 0$, and $f(x) \leq f(y)$ if $x < y$. Then*

$$\bar{s}(x,y) = \frac{s(x,y)}{f(s(x,x) + s(y,y) - s(x,y))}$$

is a similarity metric.

Proof.
Condition 1, 2, and 3: Trivial.
Condition 4: Let

$$\begin{aligned}
f_1 &= f(s(x,x) + s(y,y) + s(z,z) - s(x,y) - s(y,z)),\\
f_2 &= f(s(x,x) + s(y,y) - s(x,y)),\\
f_3 &= f(s(y,y) + s(z,z) - s(y,z)),\\
f_4 &= f(s(y,y)).
\end{aligned}$$

Consequently, $f_1 \geq \{f_2, f_3\} \geq f_4$. Further, let $a = s(y,y)$, $b = s(x,x) - s(x,y)$, and $c = s(z,z) - s(y,z)$. Therefore,

$$\begin{aligned}
\bar{s}(x,y) + \bar{s}(y,z) - \bar{s}(y,y) - \bar{s}(x,z) &= \bar{s}(x,y) + \bar{s}(y,z) - \bar{s}(y,y) - \frac{s(x,z)}{f(s(x,x) + s(z,z) - s(x,z))}\\
&\leq \bar{s}(x,y) + \bar{s}(y,z) - \bar{s}(y,y)\\
&\quad - \frac{s(x,y) + s(y,z) - s(y,y)}{f(s(x,x) + s(z,z) - (s(x,y) + s(y,z) - s(y,y)))} \qquad (2)\\
&= \frac{1}{f_1} \times \left(\frac{s(x,y) \times (f_1 - f_2)}{f_2} + \frac{s(y,z) \times (f_1 - f_3)}{f_3} - \frac{s(y,y) \times (f_1 - f_4)}{f_4}\right)\\
&\leq \frac{s(y,y)}{f_1 \times f_4} \times (f_1 + f_4 - f_2 - f_3)\\
&= \frac{s(y,y)}{f_1 \times f_4} \times (f(a+b+c) + f(a) - f(a+b) - f(a+c))\\
&\leq 0. \qquad (3)
\end{aligned}$$

The inequality in (2) clearly holds for $s(x,z) \geq 0$. When $s(x,z) < 0$, this relation also holds due to Lemma 6. The last inequality in (3) holds due to Lemma 4.
Condition 5: If $x = y$, clearly $\bar{s}(x,x) = \bar{s}(y,y) = \bar{s}(x,y)$. Conversely, if $\bar{s}(x,x) = \bar{s}(y,y) = \bar{s}(x,y)$, then $\frac{s(x,x)}{f(s(x,x))} = \frac{s(y,y)}{f(s(y,y))} = \frac{s(x,y)}{f(s(x,x)+s(y,y)-s(x,y))}$. Since $s(y,y) \geq s(x,y)$ and $f(s(y,y)) \leq f(s(x,x) + s(y,y) - s(x,y))$, in order for $\frac{s(y,y)}{f(s(y,y))} = \frac{s(x,y)}{f(s(x,x)+s(y,y)-s(x,y))}$ to hold we must have $s(y,y) = s(x,y)$. Similarly we must have $s(x,x) = s(x,y)$. This means $s(x,x) = s(y,y) = s(x,y)$, hence $x = y$. □

Theorem 2. *Let f be a function satisfying $f(0) \geq 0$, $f(x) > 0$ if $x > 0$, and $f(x) \leq f(y)$ if $x < y$. Then, given a similarity metric $s(x,y) \geq 0$,*

$$\bar{s}(x,y) = \frac{s(x,y)}{f(\max\{s(x,x), s(y,y)\})}$$

is a similarity metric.

Proof.
Condition 1, 2, and 3: Trivial.
Condition 4: To show $\bar{s}(x,y)+\bar{s}(y,z)\le\bar{s}(x,z)+\bar{s}(y,y)$, there are three cases to consider.

1. $s(z,z)\le s(x,x)\le s(y,y)$:

$$\begin{aligned}\bar{s}(x,y)+\bar{s}(y,z) &= \frac{s(x,y)}{f(s(y,y))}+\frac{s(y,z)}{f(s(y,y))}\\ &\le \frac{s(x,z)}{f(s(y,y))}+\frac{s(y,y)}{f(s(y,y))}\\ &\le \frac{s(x,z)}{f(s(x,x))}+\frac{s(y,y)}{f(s(y,y))}\\ &= \bar{s}(x,z)+\bar{s}(y,y).\end{aligned}$$

2. $s(z,z)\le s(y,y)\le s(x,x)$:

$$\begin{aligned}\bar{s}(x,y)+\bar{s}(y,z) &= \frac{f(s(y,y))\times(s(x,y)+s(y,z))+(f(s(x,x))-f(s(y,y)))\times s(y,z)}{f(s(x,x))\times f(s(y,y))}\\ &\le \frac{f(s(y,y))\times(s(x,z)+s(y,y))+(f(s(x,x))-f(s(y,y)))\times s(y,y)}{f(s(x,x))\times f(s(y,y))}\\ &= \bar{s}(x,z)+\bar{s}(y,y).\end{aligned}$$

3. $s(y,y)\le s(z,z)\le s(x,x)$:

$$\begin{aligned}\bar{s}(x,y)+\bar{s}(y,z) &= \frac{f(s(z,z))\times(s(x,y)+s(y,z))+(f(s(x,x))-f(s(z,z)))\times s(y,z)}{f(s(x,x))\times f(s(z,z))}\\ &\le \frac{f(s(z,z))\times(s(x,z)+s(y,y))+(f(s(x,x))-f(s(z,z)))\times s(y,y)}{f(s(x,x))\times f(s(z,z))}\\ &\le \frac{s(x,z)}{f(s(x,x))}+\frac{s(y,y)}{f(s(y,y))}\\ &= \bar{s}(x,z)+\bar{s}(y,y).\end{aligned}$$

Condition 5: It is clear that $\bar{s}(x,x)=\bar{s}(y,y)=\bar{s}(x,y)$ if $x=y$. Conversely, if $\bar{s}(x,x)=\bar{s}(y,y)=\bar{s}(x,y)$, then $\frac{s(x,x)}{f(s(x,x))}=\frac{s(y,y)}{f(s(y,y))}=\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\})}$. Since $s(y,y)\ge s(x,y)$ and $f(s(y,y))\le f(\max\{s(x,x),s(y,y)\})$, in order for $\frac{s(y,y)}{f(s(y,y))}=\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\})}$ to hold we must have $s(y,y)=s(x,y)$. Similarly we must have $s(x,x)=s(x,y)$. This means $s(x,x)=s(y,y)=s(x,y)$, hence $x=y$. □

Theorem 3. *Let f be a concave function over $[0,\infty)$ satisfying $f(0)\ge 0$, $f(x)>0$ if $x>0$, and $f(x)\le f(y)$ if $x<y$. Then, given a similarity metric $s(x,y)\ge 0$, for $0\le k\le 1$,*

$$\bar{s}(x,y)=\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\}+k\times(\min\{s(x,x),s(y,y)\}-s(x,y)))}$$

is a similarity metric.

Proof. We only prove that $\bar{s}(x,y)+\bar{s}(y,z)\le\bar{s}(y,y)+\bar{s}(x,z)$ as the rest is similar to the above theorems. Let

$$\begin{aligned}f_1 &= f(\max\{s(x,x),s(z,z)\}+k\times(\min\{s(x,x),s(z,z)\}-s(x,y)-s(y,z)+s(y,y))),\\ f_1' &= f(\max\{s(x,x),s(z,z)\}+k\times(\min\{s(x,x),s(z,z)\}-s(x,z))),\\ f_2 &= f(\max\{s(x,x),s(y,y)\}+k\times(\min\{s(x,x),s(y,y)\}-s(x,y))),\\ f_3 &= f(\max\{s(y,y),s(z,z)\}+k\times(\min\{s(y,y),s(z,z)\}-s(y,z))),\\ f_4 &= f(s(y,y)).\end{aligned}$$

It is straightforward to verify that all the above terms are non-negative. As will be clear soon, it is useful to sort out the relative magnitudes for $\{f_1,f_1'\}$ and $\{f_1,f_2,f_3,f_4\}$. For $\{f_1,f_1'\}$, we have $f_1\ge f_1'$ since $s(x,y)+s(y,z)\le s(x,z)+s(y,y)$. Using the fact that $s(x,z)\ge 0$, we have

$$\begin{aligned}\bar{s}(x,y)+\bar{s}(y,z)-\bar{s}(y,y)-\bar{s}(x,z) &= \frac{s(x,y)}{f_2}+\frac{s(y,z)}{f_3}-\frac{s(y,y)}{f_4}-\frac{s(x,z)}{f_1'}\\ &\le \frac{s(x,y)}{f_2}+\frac{s(y,z)}{f_3}-\frac{s(y,y)}{f_4}-\frac{s(x,z)}{f_1}\\ &\le \frac{s(x,y)\times(f_1-f_2)}{f_1\times f_2}+\frac{s(y,z)\times(f_1-f_3)}{f_1\times f_3}-\frac{s(y,y)\times(f_1-f_4)}{f_1\times f_4}.\end{aligned}$$

Table 1
A comparison of metric conditions.

Formula	k	$s(x, y)$	f
$\frac{s(x,y)}{f(s(x,x)+s(y,y)-s(x,y))}$	1	Any	Concave
$\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\}+k\times(\min\{s(x,x),s(y,y)\}-s(x,y)))}$	$0 < k < 1$	≥ 0	Concave
$\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\})}$	0	≥ 0	Any

For $\{f_1, f_2, f_3, f_4\}$, we have $\{f_2, f_3\} \geq f_4$. The full order for $\{f_1, f_2, f_3, f_4\}$ depends on the relative magnitudes of $\{s(x, x), s(y, y), s(z, z)\}$. Since x and z are symmetric in the formula, we can assume that $s(x, x) \geq s(z, z)$. Therefore there are three cases to consider, namely $s(x, x) \geq s(z, z) \geq s(y, y)$, $s(x, x) \geq s(y, y) \geq s(z, z)$, and $s(y, y) \geq s(x, x) \geq s(z, z)$. The cases $s(x, x) \geq s(z, z) \geq s(y, y)$ and $s(x, x) \geq s(y, y) \geq s(z, z)$ give rise to the partial order: $f_1 \geq \{f_2, f_3\} \geq f_4$. The case $s(y, y) \geq s(x, x) \geq s(z, z)$ results in multiple possibilities: $f_1 \geq \{f_2, f_3\} \geq f_4$, $f_2 \geq f_1 \geq f_3 \geq f_4$, $f_3 \geq f_1 \geq f_2 \geq f_4$, $\{f_2, f_3\} \geq f_1 \geq f_4$, and $\{f_2, f_3\} \geq f_4 \geq f_1$. We first derive the following result for $f_1 \geq \{f_2, f_3\} \geq f_4$ as it is relevant in all three cases:

$$\begin{aligned}\bar{s}(x, y) + \bar{s}(y, z) - \bar{s}(y, y) - \bar{s}(x, z) &\leq \frac{s(x, y) \times (f_1 - f_2)}{f_1 \times f_2} + \frac{s(y, z) \times (f_1 - f_3)}{f_1 \times f_3} - \frac{s(y, y) \times (f_1 - f_4)}{f_1 \times f_4} \\ &\leq \frac{s(y, y)}{f_1 \times f_4} \times (f_1 + f_4 - f_2 - f_3).\end{aligned}$$

Since $\frac{s(y,y)}{f_1 \times f_4} \geq 0$, in the following when $f_1 \geq \{f_2, f_3\} \geq f_4$, it suffices to prove $f_1 + f_4 - f_2 - f_3 \leq 0$.

1. $s(x, x) \geq s(z, z) \geq s(y, y)$: Let $a = s(y, y)$, $b = s(x, x) - s(y, y) + k \times (s(y, y) - s(x, y))$, $c = s(z, z) - s(y, y) + k \times (s(y, y) - s(y, z))$, and $c' = k \times (s(z, z) - s(y, z))$. From $c - c' = (1 - k) \times (s(z, z) - s(y, y)) \geq 0$, we have $c \geq c'$. Then, $f_1 + f_4 - f_2 - f_3 = f(a + b + c') + f(a) - f(a + b) - f(a + c) \leq f(a + b + c) + f(a) - f(a + b) - f(a + c) \leq 0$.
2. $s(x, x) \geq s(y, y) \geq s(z, z)$: Let $a = s(y, y)$, $b = s(x, x) - s(y, y) + k \times (s(y, y) - s(x, y))$, and $c = k \times (s(z, z) - s(y, z))$. Then, $f_1 + f_4 - f_2 - f_3 = f(a + b + c) + f(a) - f(a + b) - f(a + c) \leq 0$.
3. $s(y, y) \geq s(x, x) \geq s(z, z)$:
   - $f_1 \geq \{f_2, f_3\} \geq f_4$: Similar as above.
   - $f_2 \geq f_1 \geq f_3 \geq f_4$: Using the fact that $s(y, y) \geq s(y, z)$, we have

$$\begin{aligned}\bar{s}(x, y) + \bar{s}(y, z) - \bar{s}(y, y) - \bar{s}(x, z) &\leq \frac{s(x, y) \times (f_1 - f_2)}{f_1 \times f_2} + \frac{s(y, z) \times (f_1 - f_3)}{f_1 \times f_3} - \frac{s(y, y) \times (f_1 - f_4)}{f_1 \times f_4} \\ &\leq \frac{s(x, y) \times (f_1 - f_2)}{f_1 \times f_2} + \frac{s(y, y) \times (f_1 - f_4)}{f_1 \times f_3 \times f_4} \times (f_4 - f_3) \\ &\leq 0.\end{aligned}$$

   - $f_3 \geq f_1 \geq f_2 \geq f_4$: Similar as above.
   - $\{f_2, f_3\} \geq f_1 \geq f_4$ or $\{f_2, f_3\} \geq f_4 \geq f_1$: Using the fact that $\min\{f_2, f_3\} \geq \max\{f_1, f_4\}$, we have

$$\begin{aligned}\bar{s}(x, y) + \bar{s}(y, z) - \bar{s}(y, y) - \bar{s}(x, z) &\leq \frac{s(x, y)}{f_2} + \frac{s(y, z)}{f_3} - \frac{s(y, y)}{f_4} - \frac{s(x, z)}{f_1} \\ &\leq \frac{s(x, y) + s(y, z) - s(y, y) - s(x, z)}{\max\{f_1, f_4\}} \\ &\leq 0.\end{aligned}$$

Therefore, we have proved that $\bar{s}(x, y) + \bar{s}(y, z) \leq \bar{s}(y, y) + \bar{s}(x, z)$. □

Note. We may define $\bar{s}(x, y) = 0$ if both the numerator and the denominator are 0.

Corollary 3. *In the above theorems, we obtain normalized similarity metrics with an additional condition, $f(x) \geq x$.*

Proof. Trivial. □

Remark. We see that $\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\}+k\times(\min\{s(x,x),s(y,y)\}-s(x,y)))}$ can reduce to $\frac{s(x,y)}{f(s(x,x)+s(y,y)-s(x,y))}$ or $\frac{s(x,y)}{f(\max\{s(x,x),s(y,y)\})}$ with $k = 1$ or $k = 0$, respectively. A comparison of their respective metric conditions is listed in Table 1. When k is in between 0 and 1 the condition requirement is more stringent than when k takes on the limits, i.e. 0 or 1. When $k = 0$ the condition for f is relaxed, whereas when $k = 1$ the condition for $s(x, y)$ is relaxed.

5. Normalized distance metric

Theorem 4. *Let $d(x, y)$ be a distance metric on X. Let f be a concave function on $[0, \infty)$ such that $f(0) \geq 0$, $f(x) > 0$ if $x > 0$, and $f(x) \leq f(y)$ if $x < y$. Then for any fixed $o \in X$,*

$$\bar{d}(x, y) = \frac{d(x, y)}{f(d(x, y) + \frac{d(x,o)+d(y,o)}{k})}$$

is a distance metric, where $k \geq 1$.

Proof. We prove that $\bar{d}(x, y) \leq \bar{d}(x, z) + \bar{d}(y, z)$ as the rest is trivial.

$$\begin{aligned} \bar{d}(x, y) &= \frac{d(x, y)}{f(d(x, y) + \frac{d(x,o)+d(y,o)}{k})} \\ &\leq \frac{d(x, z) + d(y, z)}{f(d(x, z) + d(y, z) + \frac{d(x,o)+d(y,o)}{k})} \\ &\leq \frac{d(x, z)}{f(d(x, z) + \frac{d(x,o)+d(z,o)}{k})} + \frac{d(y, z)}{f(d(y, z) + \frac{d(y,o)+d(z,o)}{k})} \\ &= \bar{d}(x, z) + \bar{d}(y, z). \end{aligned} \tag{4}$$

The inequality in (4) is due to Lemma 6. □

Corollary 4. *With an additional condition that $f(x) \geq \frac{k}{k+1} \times x$,*

$$\frac{d(x, y)}{f(d(x, y) + \frac{d(x,o)+d(y,o)}{k})}$$

is a normalized distance metric.

Proof. Trivial. □

Theorem 5. *Let $d(x, y)$ be a distance metric on X. Let f be a function such that $f(0) \geq 0$, $f(x) > 0$ if $x > 0$, and $f(x) \leq f(y)$ if $x < y$. Then for any fixed $o \in X$,*

$$\bar{d}(x, y) = \frac{d(x, y) - \min\{d(x, o), d(y, o)\}}{f(\max\{d(x, o), d(y, o)\})} + \frac{\min\{d(x, o), d(y, o)\}}{f(\min\{d(x, o), d(y, o)\})}$$

is a distance metric.

Proof. Let $s(x, y) = d(x, o) + d(y, o) - d(x, y)$, then from Lemma 10, $s(x, y)$ is a non-negative similarity metric. Since $f(\frac{x}{2})$ satisfies the conditions of Theorem 2, the following is a similarity metric

$$\bar{s}(x, y) = \frac{d(x, o) + d(y, o) - d(x, y)}{f(\max\{d(x, o), d(y, o)\})}.$$

Applying Lemma 8 to $\bar{s}(x, y)$, we have that

$$\frac{d(x, o)}{f(d(x, o))} + \frac{d(y, o)}{f(d(y, o))} - \frac{d(x, o) + d(y, o) - d(x, y)}{f(\max\{d(x, o), d(y, o)\})}$$

is a distance metric. Therefore

$$\bar{d}(x, y) = \frac{d(x, y) - \min\{d(x, o), d(y, o)\}}{f(\max\{d(x, o), d(y, o)\})} + \frac{\min\{d(x, o), d(y, o)\}}{f(\min\{d(x, o), d(y, o)\})}$$

is a distance metric.

Note that from the formula, $\bar{d}(x, o)$ needs special definition. We can define $\bar{d}(o, o) = 0$ and $\bar{d}(x, o) = \frac{d(x,o)}{f(d(x,o))}$. □

Corollary 5. *With an additional condition that $f(x) \geq 2 \times x$,*

$$\frac{d(x, y) - \min\{d(x, o), d(y, o)\}}{f(\max\{d(x, o), d(y, o)\})} + \frac{\min\{d(x, o), d(y, o)\}}{f(\min\{d(x, o), d(y, o)\})}$$

is a normalized distance metric.

Proof. Trivial. □

Table 2
Summary: Set similarity metrics and distance metrics.

Similarity	Distance
$\lvert A \cap B\rvert$	$\lvert A-B\rvert + \lvert B-A\rvert$
	$\max\{\lvert A-B\rvert, \lvert B-A\rvert\}$
$\frac{\lvert A\cap B\rvert}{\lvert A\cup B\rvert}$	$\frac{\lvert A-B\rvert+\lvert B-A\rvert}{\lvert A\cup B\rvert}$
$\frac{\lvert A\cap B\rvert}{\max\{\lvert A\rvert,\lvert B\rvert\}}$	$\frac{\max\{\lvert A-B\rvert,\lvert B-A\rvert\}}{\max\{\lvert A\rvert,\lvert B\rvert\}}$

Corollary 6. *With an additional condition that $f(x)$ is concave,*

$$\frac{d(x,y) - \min\{d(x,o), d(y,o)\} + \max\{d(x,o), d(y,o)\}}{f(\max\{d(x,o), d(y,o)\})}$$

is a distance metric.

Proof. In the last step of the proof of the theorem, applying Lemma 9 instead of Lemma 8 and using the following fact

$$\max\left\{\frac{d(x,o)}{f(d(x,o))}, \frac{d(y,o)}{f(d(y,o))}\right\} = \frac{\max\{d(x,o), d(y,o)\}}{f(\max\{d(x,o), d(y,o)\})},$$

the result follows. □

6. Examples

In specific problem settings, several similarity and distance metrics have been proposed, for example, in finding maximal common subgraph between two graphs, in defining information distance based on the notion of Kolmogorov complexity, and in evaluating the importance of attributes. These are special solutions, each of which is only suitable for a specific context from which it is derived. In the following, we show that by casting the solutions in previous sections to each of these contexts, these metrics readily follow.

6.1. Set similarity and distance

Given sets A and B, we denote by $A - B$ the relative complement of B in A, i.e. $A - B = A \cap \bar{B} = \{x \in A \mid x \notin B\}$.

Graph distance: An example of graph distance metric [2], based on the notion of maximal common subgraph, is $1 - \frac{|G_1 \cap G_2|}{\max\{|G_1|,|G_2|\}} = \frac{\max\{|G_1-G_2|,|G_2-G_1|\}}{\max\{|G_1|,|G_2|\}}$ where $G_1 \cap G_2$ represents the maximal common subgraph between the graphs G_1 and G_2 and $|G_1 \cap G_2|$ is a similarity metric.

Attributed tree distance: An attributed tree is a tree of which every node is associated with a vector of attributes. A way of defining a distance metric between two attributed trees is based on *maximum similarity subtree isomorphism* [24]. Examples are

- $|T_1| + |T_2| - 2 \times |T_1 \cap T_2| = |T_1 - T_2| + |T_2 - T_1|$,
- $\max\{|T_1|, |T_2|\} - |T_1 \cap T_2| = \max\{|T_1 - T_2|, |T_2 - T_1|\}$,
- $1 - \frac{|T_1\cap T_2|}{|T_1 \cup T_2|} = \frac{|T_1-T_2|+|T_2-T_1|}{|T_1\cup T_2|}$,
- $1 - \frac{|T_1\cap T_2|}{\max\{|T_1|,|T_2|\}} = \frac{\max\{|T_1-T_2|,|T_2-T_1|\}}{\max\{|T_1|,|T_2|\}}$

where $T_1 \cap T_2$ represents a maximum similarity subtree between two attributed trees T_1 and T_2 and $|T_1 \cap T_2|$ is a similarity metric.

The formulation of the metrics in the above examples is essentially based on the notion of set similarity and distance. Therefore, we now cast the general solution in this context.

Given finite sets A, B and C, we have $|A \cap B| + |B \cap C| - |A \cap C| \le |B|$. Note that this inequality is the equivalent of that in similarity condition 4. It is easy to verify that $|A \cap B|$ is a similarity metric. From Lemmas 8 and 9 it follows that both $|A - B| + |B - A|$ and $\max\{|A - B|, |B - A|\}$ are distance metrics. From Theorem 1, it follows that $\frac{|A\cap B|}{|A\cup B|}$ is a similarity metric and consequently $\frac{|A-B|+|B-A|}{|A\cup B|}$ is a distance metric. From Theorem 2, it follows that $\frac{|A\cap B|}{\max\{|A|,|B|\}}$ is a similarity metric and consequently $\frac{\max\{|A-B|,|B-A|\}}{\max\{|A|,|B|\}}$ is a distance metric.

We summarize the results in Table 2.

Remark. Note that these are a subset of the metrics that may derive from the general solution. Evidently they encompass the metrics in the examples. For the formulae in fractional forms, we have chosen a simple concave function $f(x) = x$. There are many other functions to choose, so long as they meet the functional conditions specified in previous sections. It is, in general, easier to determine whether a given function meets a set of conditions than to prove that a given formula involving this function is a metric.

Table 3
Summary: Similarity and distance metrics for evaluating the importance of attributes.

Similarity	Distance
$I(X,Y)$, [6,9,16]	$H(X\|Y)+H(Y\|X)$, [10,25]
$\frac{I(X,Y)}{H(X,Y)}$, [12]	$\frac{H(X\|Y)+H(Y\|X)}{H(X,Y)}$, [10,17]
$\frac{I(X,Y)}{\max\{H(X),H(Y)\}}$, [5,7]	

Table 4
Summary: Information similarity metrics and distance metrics.

Similarity	Distance
$I(X,Y)$	$H(X\|Y)+H(Y\|X)$
	$\max\{H(X\|Y),H(Y\|X)\}$
$\frac{I(X,Y)}{H(X,Y)}$	$\frac{H(X\|Y)+H(Y\|X)}{H(X,Y)}$
$\frac{I(X,Y)}{\max\{H(X),H(Y)\}}$	$\frac{\max\{H(X\|Y),H(Y\|X)\}}{\max\{H(X),H(Y)\}}$

6.2. *Information similarity and distance*

Kolmogorov complexity: There has been some study on defining the information distance, based on the notion of Kolmogorov complexity [8]. The Kolmogorov complexity $K(x)$ of a string x is the length of a shortest binary program x^* to compute x on an appropriate universal computer. The distance between two objects may be defined to be the length of the shortest program that can transform either object into the other and vice versa. Examples of such information distance metric are $\frac{K(x|y^*)+K(y|x^*)}{K(x,y)}$ and $\frac{\max\{K(x|y^*),K(y|x^*)\}}{\max\{K(x),K(y)\}}$.
Data mining: An attribute is deemed important in data mining if it partitions the database such that new patterns are revealed [27]. Several similarity and distance metrics were proposed in the context of evaluating the importance of attributes. They are listed in Table 3.

The formulation of the metrics in the above examples is essentially based on the notion of information similarity and distance. We now cast the general solution in this context.

Denote by $H(X)$ the information entropy of a discrete random variable X, $H(Y|X)$ the entropy of Y conditional on X, $H(X,Y)$ the joint entropy of X and Y, and $I(X,Y)$ the mutual information between X and Y.

From information theory, we have $H(X|Y) \leq H(X|Z) + H(Z|Y)$. The mutual information between X and Y is defined as $I(X,Y) = H(X) - H(X|Y)$, with $I(X,Y) = I(Y,X)$. With the above, we have $I(X,Y) + I(Y,Z) \leq I(X,Z) + I(Y,Y)$. Then, it is straightforward to verify that $I(X,Y)$ is a similarity metric. From Lemmas 8 and 9 it follows that both $H(X|Y)+H(Y|X)$ and $\max\{H(X|Y), H(Y|X)\}$ are distance metrics. From Theorem 1, it follows that $\frac{I(X,Y)}{H(X,Y)}$ is a similarity metric where $H(X,Y)$ is the joint entropy of X and Y defined as $H(X,Y) = H(X) + H(Y|X)$ with $H(X,Y) = H(Y,X)$. Consequently, $\frac{H(X|Y)+H(Y|X)}{H(X,Y)}$ is a distance metric. From Theorem 2, it follows that $\frac{I(X,Y)}{\max\{H(X),H(Y)\}}$ is a similarity metric. Consequently, $\frac{\max\{H(X|Y),H(Y|X)\}}{\max\{H(X),H(Y)\}}$ is a distance metric.

We summarize the results in Table 4.

Remark. Note the resemblance between the above metrics and those for the case of set, both constructed from the general solution. Furthermore, it is evident that the metrics in these examples can all be obtained from the same principle. In the context of Kolmogorov complexity, basic quantities such as $K(x)$, $K(x,y)$, $K(x|y)$ and $I(x:y)$ are similar to $H(X)$, $H(X,Y)$, $H(X|Y)$ and $I(X,Y)$, respectively. Their respective formulae take on equivalent forms. Analogous to $I(X,Y)$, $I(x:y)$ is a similarity metric. With this, the two distance metrics readily follow from the general solution.

6.3. *Sequence edit distance and similarity*

It is well known that if the cost for basic operations of insertion, deletion, and substitution is a distance metric, then the sequence edit distance $d(s_1,s_2)$, defined between two sequences s_1 and s_2 and derived from finding the minimum-cost sequence of operations that transform s_1 to s_2, is also a distance metric.

Several normalized edit distances have been proposed and studied [13,15]. Examples are $\frac{d(s_1,s_2)}{|s_1|+|s_2|}$, $\frac{d(s_1,s_2)}{\max\{|s_1|,|s_2|\}}$, and $n(s_1,s_2) = \min\{\frac{d(s_1,s_2)}{|p|} \mid p \text{ is a path that changes } s_1 \text{ to } s_2\}$. Although these are referred to as normalized edit distance, they are not distance metric.

From the results of Section 5, choosing o as the empty sequence, we have two normalized edit distance metrics. If the indel cost is 1, then the following is a normalized distance metric:

$$\frac{1}{2} \times \left(\frac{d(s_1, s_2) - \min\{|s_1|, |s_2|\}}{\max\{|s_1|, |s_2|\}} + 1 \right).$$

For sequence similarity, one popular measurement is protein sequence similarity based on BLOSUM matrices using Smith–Waterman algorithm [20]. In fact, based on the original score without rounding, any BLOSUM-N matrix with $N \geq 55$ is a similarity metric. Therefore protein sequence similarity based on those BLOSUM matrices with Smith–Waterman algorithm is a similarity metric.

For normalized sequence similarity, an example is $\frac{s(s_1,s_2)}{|s_1|+|s_2|+k}$ where $k > 0$ [1]. This, however, is not a similarity metric since condition 4 of the similarity metric is not satisfied.

7. Conclusions

We have given a formal definition for the similarity metric. We have shown the relationship between the similarity metric and the distance metric. We have given general formulae to normalize a similarity metric or a distance metric. We have shown, with examples, how the general solutions are useful in constructing metrics suitable for various contexts.

Acknowledgments

We thank an anonymous referee for offering helpful suggestions in improving the presentation of this article. Part of the work was done while Kaizhong Zhang was visiting the Institute for Mathematical Sciences, National University of Singapore in 2006. The visit was partially supported by the Institute.

References

[1] A.N. Arslan, Oï Eğecioğlu, P.A. Pevzner, A new approach to sequence alignment: Normalized sequence alignment, Bioinformatics 17 (4) (2001) 327–337.
[2] H. Bunke, K. Shearer, A graph distance metric based on the maximal common subgraph, Pattern Recognition Letters 19 (1998) 255–259.
[3] C.S. Calude, K. Salomaa, S. Yu, Additive distances and quasi-distances between words, Journal of Universal Computer Science 8 (2) (2002) 141–152.
[4] S. Chen, B. Ma, K. Zhang, The normalized similarity metric and its applications, in: Proceedings of 2007 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2007, 2007, pp. 172–180.
[5] Y. Horibe, Entropy and correlation, IEEE Transactions on Systems, Man, and Cybernetics 15 (1985) 641–642.
[6] A.J. Knobbe, P.W. Adriaans, Analysing binary associations, in: E. Simoudis, J. Han, U. Fayyad (Eds.), Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 311–314.
[7] T.O. Kvålseth, Entropy and correlation: Some comments, IEEE Transactions on Systems, Man, and Cybernetics 17 (1987) 517–519.
[8] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitányi, The similarity metric, IEEE Transactions on Information Theory 50 (12) (2004) 3250–3264.
[9] E.H. Linfoot, An informational measure of correlation, Information and Control 1 (1) (1957) 85–89.
[10] R. López de Mántaras, Id3 revisited: A distance-based criterion for attribute selection, in: Z. Ras (Ed.), Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems, 1989, pp. 342–350.
[11] B. Ma, K. Zhang, The similarity metric and the distance metric, in: Proceedings of the 6th Atlantic Symposium on Computational Biology and Genome Informatics, 2005, pp. 1239–1242.
[12] F.M. Malvestuto, Statistical treatment of the information content of a database, Information Systems 11 (1986) 211–223.
[13] A. Marzal, E. Vidal, Computation of normalized edit distance and applications, IEEE Transactions on Pattern Analysis and Machine Intelligence 15 (9) (1993) 926–932.
[14] S.E. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino-acid sequences of two proteins, Journal of Molecular Biology 48 (1970) 443–453.
[15] B.J. Oommen, K. Zhang, The normalized string editing problem revisited, IEEE Transactions on Pattern Analysis and Machine Intelligence 18 (6) (1996) 669–672.
[16] J.R. Quinlan, Induction of decision trees, Machine Learning 1 (1) (1986) 81–106.
[17] C. Rajski, A metric space of discrete probability distributions, Information and Control 4 (4) (1961) 371–377.
[18] S.C. Sahinalp, M. Tasan, J. Macker, Z.M. Ozsoyoglu, Distance based indexing for string proximity search, in Proceedings of the 19th International Conference on Data Engineering, 2003, pp. 125–136.
[19] N. Saitou, M. Nei, The neighbor-joining method: A new method for reconstructing phylogenetic trees, Molecular Biology and Evolution 4 (1987) 406–425.
[20] T.F. Smith, M.S. Waterman, Comparison of biosequences, Advances in Applied Mathematics 2 (1981) 482–489.
[21] R.R. Sokal, C.D. Michener, A statistical method for evaluating systematic relationships, University of Kansas Scientific Bulletin 28 (1958) 1409–1438.
[22] A. Stojmirovic, V. Pestov, Indexing schemes for similarity search in datasets of short protein fragments, ArXiv Computer Science e-prints (cs/0309005), September 2003.
[23] J.A. Studier, K.J. Keppler, A note on the neighbor-joining algorithm of Saitou and Nei, Molecular Biology and Evolution 5 (1988) 729–731.
[24] A. Torsello, D. Hidović-Rowe, M. Pelillo, Polynomial-time metrics for attributed trees, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (7) (2005) 1087–1099.
[25] S.J. Wan, S.K.M. Wong, A measure for concept dissimilarity and its application in machine learning, in: Proceedings of the First International Conference on Computing and Information, 1989, pp. 267–273.
[26] M.S. Waterman, T.F. Smith, Some biological sequence metrics, Advances in Mathematics 20 (1976) 367–387.
[27] Y.Y. Yao, S.K.M. Wong, C.J. Butz, On information-theoretic measures of attribute importance, in: N. Zhong (Ed.), Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining, 1999, pp. 133–137.
[28] K. Zhang, D. Shasha, Simple fast algorithms for the editing distance between trees and related problems, SIAM Journal on Computing 18 (6) (1989) 1245–1262.