Electronic Supplementary Materials

for the paper

"A new class of metrics for learning on real-valued and structured data" by R. Yang, Y. Jiang, S. Mathews, E. A. Housworth, M. W. Hahn, and P. Radivojac. *Data Mining and Knowledge Discovery*, 2019.

This document contains all supplementary information for the paper "A new class of metrics for learning on real-valued and structured data" by Yang, Jiang *et al.* that appears in *Data Mining and Knowledge Discovery* Journal. The sections in this Supplement are numbered in continuation to the main paper in order to enable equation and theorem referencing in the main paper as well as referencing additional lemmas necessary to prove these theorems.

9. Lemmas and Proofs

Proof of Theorem 3.1. The only metric property not obviously satisfied by d is the triangle inequality. Given arbitrary sets $A, B, C \in X$, we have

$$\begin{aligned} d(A,B) + d(B,C) &= (|A \setminus B|^p + |B \setminus A|^p)^{\frac{1}{p}} + (|B \setminus C|^p + |C \setminus B|^p)^{\frac{1}{p}} \\ &\geq ((|A \setminus B| + |B \setminus C|)^p + (|B \setminus A| + |C \setminus B|)^p)^{\frac{1}{p}} \\ &\geq (|A \setminus C|^p + |C \setminus A|^p)^{\frac{1}{p}} \\ &= d(A,C). \end{aligned}$$

The first inequality holds due to Minkowski inequality. The second inequality can be deduced from triangle inequality of the Manhattan distance on binary vectors. □

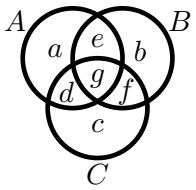

Figure 9. The Venn diagram and cardinality related to sets A, B and C.

Proof of Theorem 3.2. As in the unnormalized case we only prove the triangle inequality. Let $A, B, C \in X$ be arbitrary sets. In a special case when at least one of $|A \cup B| = 0$ and $|B \cup C| = 0$ holds, the triangle inequality holds. To avoid division by zero, we next assume all denominators are positive; i.e., we only consider the cases when $|A \cup B| \neq 0$ and $|B \cup C| \neq 0$. Let the cardinality of different sets shown in Figure 9 be denoted by a, b, ..., g. Let $\tau = |A \cup B \cup C| = a + b + \cdots + g$ and $\mathrm{h_p}(x, y) = (x^p + y^p)^{1/p}$.

$$
\begin{aligned}
d_N(A,B) + d_N(B,C) &= \frac{\mathrm{h_p}(a+d, b+f)}{\tau - c} + \frac{\mathrm{h_p}(b+e, c+d)}{\tau - a} \\
&\geq \frac{\mathrm{h_p}(a+d, b+f)}{\tau} + \frac{\mathrm{h_p}(b+e, c+d)}{\tau} \\
&\geq \frac{\mathrm{h_p}(a+d+b+e, b+f+c+d)}{\tau} \\
&= \mathrm{h_p}\left(\frac{a+d+b+e}{\tau}, \frac{b+f+c+d}{\tau}\right) \\
&\geq \mathrm{h_p}\left(\frac{a+d+b+e-b}{\tau - b}, \frac{b+f+c+d-b}{\tau - b}\right) \\
&\geq \frac{\mathrm{h_p}(a+d+e, f+c+d)}{\tau - b} \\
&\geq \frac{\mathrm{h_p}(a+e, f+c)}{\tau - b} \\
&= d_N(A,C).
\end{aligned}
$$

The second inequality holds due to Minkowski inequality. The third inequality holds since we subtracted the same nonnegative number b from both the numerator and denominator of the fraction with the fraction itself remaining in $[0,1]$ after the subtraction (the numerator is nonnegative before and after the subtraction). Hence, d_N is a metric. It follows that d_N is bounded in $[0,1]$ via the Minkowski inequality. □

Proof of Theorem 3.3 and Theorem 3.4. The metrics d and d_N defined in Eqs. 6-7 can be reduced to be applied on vectors in $\mathbb{R}^k$ by rewriting $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ as a map $f(t) := \sum x_i \mathbf{1}_{i-1<t\leq i}$, where $\mathbf{1}_{(\cdot)}$ is an indicator function. Since d and d_N are metrics, proved in Theorem 3.5 and Theorem 3.6, we obtain that d and d_N on $\mathbb{R}^k$ are also metrics. □

Lemma 9.1. *For any two real functions f and g, we have $(f+g)^+ \leq f^+ + g^+$, and, $(f+g)^- \leq f^- + g^-$.*

Proof. We will prove $(f+g)^+ \leq f^+ + g^+$ first. Since

$$f^+ + g^+ = \max(f, 0) + \max(g, 0) \geq f + g,$$

and

$$f^+ + g^+ = \max(f, 0) + \max(g, 0) \geq 0,$$

we have that $f^+ + g^+ \geq \max(f+g, 0) = (f+g)^+$.

Notice that $f^- = (-f)^+$. To prove $(f+g)^- \leq f^- + g^-$, we have $(f+g)^- = (-f-g)^+ \leq (-f)^+ + (-g)^+ = f^- + g^-$. □

Lemma 9.2. *For any real functions f, g we have*

$$f^+ + g^+ = \min(|f|, |g|)\mathbf{1}_{\{fg<0\}} + (f+g)^+.$$

Proof. For $fg \geq 0$, it is not hard to see that $f^+ + g^+ = (f+g)^+$. When $fg < 0$, without loss of generality, suppose that $f < 0$ and $g > 0$. Then we have

$$\begin{aligned} f^+ + g^+ - (f+g)^+ &= g - (f+g)^+ \\ &= \begin{cases} g, & \text{if } |f| > |g| \\ -f, & \text{if } |f| \leq |g| \end{cases} = \min(|f|, |g|). \end{aligned}$$

□

Lemma 9.3. *For any real functions f, g and h, we have*

$$(f-g)^+ + (g-h)^+ - \{\max(|g|, |f-g|, |g-h|) - \max(|f|, |h|, |f-h|)\}^+ \geq (f-h)^+.$$

Proof. By Lemma 9.2, it is equivalent to show

$$\min(|f-g|, |g-h|)\mathbf{1}_{\{(f-g)(g-h)<0\}} \geq \{\max(|g|, |f-g|, |g-h|) - \max(|f|, |h|, |f-h|)\}^+.$$

Since $\min(|f-g|, |g-h|)\mathbf{1}_{\{(f-g)(g-h)<0\}} \geq 0$, it suffices to show that it is no less than $\max(|g|, |f-g|, |g-h|) - \max(|f|, |h|, |f-h|)$.

Consider the case $(f-g)(g-h) \geq 0$ first. We have $f \geq g \geq h$ or $f \leq g \leq h$. g is located between f and h, therefore $|g| \leq \max(|f|, |h|)$ and $\max(|f-g|, |g-h|) \leq |f-h|$. Then we have

$$|g| - \max(|f|, |h|, |f-h|) \leq |g| - \max(|f|, |h|) \leq 0,$$

and

$$\max(|f-g|, |g-h|) - \max(|f|, |h|, |f-h|) \leq \max(|f-g|, |g-h|) - |f-h| \leq 0.$$

This shows that

$$\max(|g|.|f-g|, |g-h|) - \max(|f|, |h|, |f-h|) \leq\ 0 = \min(|f-g|, |g-h|)\mathbf{1}_{\{(f-g)(g-h)<0\}}$$

When $(f-g)(g-h) \leq 0$, we have the following two cases

(1) $\min(|f-g|, |g-h|) = |f-g|$
(2) $\min(|f-g|, |g-h|) = |g-h|$

For case 1, we want to show that $|f-g| \geq \max(|g|, |g-h|) - \max(|f|, |h|, |f-h|)$. This is true since

$$|f-g| \geq |g| - |f| \geq |g| - \max(|f|, |h|, |f-h|),$$

and

$$|f-g| \geq |g-h| - |f-h| \geq |g-h| - \max(|f|, |h|, |f-h|).$$

Combining those two inequalities we get $|f-g| \geq \max(|g|, |g-h|) - \max(|f|, |h|, |f-h|)$. For case 2, we want to show $|g-h| \geq \max(|g|, |f-g|) - \max(|f|, |h|, |f-h|)$. By the same analogy, since

$$|g-h| \geq |g| - |h| \geq |g| - \max(|f|, |h|, |f-h|),$$

and

$$|g-h| \geq |f-g| - |f-h| \geq |f-g| - \max(|f|, |h|, |f-h|),$$

we have $|g-h| \geq \max(|g|, |g-h|) - \max(|f|, |h|, |f-h|)$. Therefore this inequality still holds for $(f-g)(g-h) \leq 0$. □

Lemma 9.4. *For any two real functions f and g, we have*

$$\max(f, g) = f + (g - f)^+.$$

Lemma 9.5. *Any Cauchy sequence $\{f_n\}$ in $(L(\mathbb{R}), d_N)$ is bounded in L^1; i.e., $\int |f_n|\, dx < M$ for some constant $M > 0$.*

Proof. We instead prove the contrapositive version of the above statement. Suppose $\{f_n\}$ is a sequence in $(L(\mathbb{R}), d_N)$ that is unbounded in L^1, or equivalently $\int |f_n|\, dx = \infty$ as $n \to \infty$. Thus, we have

$$\begin{aligned} d_N(f_n, f_m) =& \frac{((\int (f_n - f_m)^-\, dx)^p + (\int (f_n - f_m)^+\, dx)^p)^{\frac{1}{p}}}{\int \max(|f_n|, |f_m|, |f_m - f_n|)\, dx} \\ \geq& \frac{((\int (f_n - f_m)^-\, dx)^p + (\int (f_n - f_m)^+\, dx)^p)^{\frac{1}{p}}}{\int |f_m| + |f_n|)\, dx} \end{aligned}$$

For any integer $N_0 > 0$, pick n to be N_0 and we have that $d_N(f_{N_0}, f_m)$ is given by

$$\begin{aligned} & \frac{((\int (f_{N_0} - f_m)^-\, dx)^p + (\int (f_{N_0} - f_m)^+\, dx)^p)^{\frac{1}{p}}}{\int |f_m| + |f_{N_0}|)\, dx} \\ =& ((\int (A_m - B_m)^-\, dx)^p + (\int (A_m - B_m)^+\, dx)^p)^{\frac{1}{p}} \end{aligned}$$

where

$$A_m = \frac{f_{N_0}}{\int |f_m| + |f_{N_0}|)\, dx}, \qquad B_m = \frac{f_m}{\int |f_m| + |f_{N_0}|)\, dx}$$

and $\{A_m\}$ and $\{B_m\}$ are sequences of functions. Fixing N_0 and sending n to ∞ we have that $\int |A_m|\, dx \to 0$ and $\int |B_m|\, dx \to 1$ as $m \to \infty$. Therefore, we could choose $m > N_0$ such that $\int |A_m|\, dx < 1/10$ and $\int |B_m|\, dx > 9/10$. Then between $\int B_m^-\, dx$ and $\int B_m^+\, dx$ there is at least one greater than $9/20$. Without loss of generality, suppose $\int B_m^-\, dx > 9/20$, then

$$\begin{aligned} d_N(f_{N_0}, f_m) &\geq (\int (A_m - B_m)^+\, dx)^p)^{\frac{1}{p}} \\ &= \int (B_m - A_m)^-\, dx \\ &\geq \int (B_m)^-\, dx - \int (A_m)^-\, dx \\ &\geq \int (B_m)^-\, dx - \int |A_m|\, dx \\ &> \frac{7}{20} \end{aligned}$$

With this we have shown that for any $N_0 > 0$, there exist $m, n \geq N_0$ such that $d_N(f_n, f_m) > 7/20$; i.e., $\{f_n\}$ is not a Cauchy sequence in $(L(\mathbb{R}), d_N)$. Thus, we have proved the claim. □

Proof of Theorem 6.1. To show that d is a metric is analogous to the proof of Theorem 3.1. Let A, B, C be arbitrary consistent subgraphs of the ontology. Instead of $|A \setminus B|$, we use $\sum_{v \in A \setminus B} ia(v)$ and similarly for the other cardinalities. The proof follows line for line after these substitutions to the proof of Theorem 3.1.

To prove d_N in Theorem 6.2 is a metric, we follow a similar argument to the proof of Theorem 3.2. The analogues of a, b, c, d, e, f, g here are

$$a = \sum_{v \in A \setminus (B \cup C)} ia(v), \quad b = \sum_{v \in B \setminus (A \cup C)} ia(v), \quad c = \sum_{v \in C \setminus (A \cup B)} ia(v), \quad d = \sum_{v \in A \cap C \setminus B} ia(v),$$
$$e = \sum_{v \in A \cap B \setminus C} ia(v), \quad f = \sum_{v \in B \cap C \setminus A} ia(v), \quad g = \sum_{v \in A \cap B \cap C} ia(v).$$

With these substitutions, the proof is exactly the same as that of Theorem 3.2. Invoking the Minkowski inequality, we obtain that

$$d_N(F, G) \leq \frac{ru(F, G) + mi(F, G)}{\sum_{v \in F \cup G} ia(v)} \leq \frac{\sum_{v \in F \cup G} ia(v)}{\sum_{v \in F \cup G} ia(v)} = 1\,.$$

Since d_N is nonnegative, we obtain that $d_N \in [0, 1]$. □

Proof of Theorem 3.5. Since d_N is non-negative, the equations $d_N(f, g) = d_N(g, f)$ and $d_N(f, g) = 0$ hold if and only if $f = g$ almost everywhere, it suffices to show that d satisfies the triangle inequality. Let f, g, and h be in $L(\mathbb{R})$. Then we have

$$\begin{aligned}
&D(f, g) + D(g, h) \\
&= \left(\left(\int (f(x) - g(x))^+ dx\right)^p + \left(\int (f(x) - g(x))^- dx\right)^p\right)^{\frac{1}{p}} \\
&+ \left(\left(\int (g(x) - h(x))^+ dx\right)^p + \left(\int (g(x) - h(x))^- dx\right)^p\right)^{\frac{1}{p}} \\
&\geq \left(\left(\int (f(x) - g(x))^+ + (g(x) - h(x))^+ dx\right)^p + \left(\int (f(x) - g(x))^- + (g(x) - h(x))^- dx\right)^p\right)^{\frac{1}{p}} \\
&\geq \left(\left(\int (f(x) - g(x) + g(x) - h(x))^+ dx\right)^p + \left(\int (f(x) - g(x) + g(x) - h(x))^- dx\right)^p\right)^{\frac{1}{p}} \\
&\geq \left(\left(\int (f(x) - h(x))^+ dx\right)^p + \left(\int (f(x) - h(x))^- dx\right)^p\right)^{\frac{1}{p}} \\
&= D(f, h).
\end{aligned}$$

Therefore, d is a metric. □

Proof of Theorem 3.6. It is easy to check that d_N is non-negative, $d_N(f, g) = d_N(g, f)$ and $d_N(f, g) = 0$ if and only if $f = g$ almost everywhere. Therefore, it remains to be shown that the inequality $d_N(f, g) + d_N(g, h) \geq d_N(f, h)$ is satisfied.

Let f, g, and h be bounded functions in $L(\mathbb{R})$. To begin, let us look at the trivial cases. Define $\mathbf{M}(f, g) = \int \max(|f|, |g|, |f - g|)\, dx$ and $\mathbf{M}^*(f, g, h) = \int \max(|f|, |g|, |h|, |f - g|, |g - h|, |f - h|)\, dx$.

a. If $\int \max(|f|, |g|, |f - g|)\, dx = 0$, then $f = g$ almost everywhere. Consequently $d_N(f, h) = d_N(g, h)$ and $d_N(f, g) = 0$, so the inequality holds.
b. If $\int \max(|f|, |h|, |f - h|)\, dx = 0$, then $d_N(f, h) = 0$, in which case the inequality is true due to the non-negativity of d_N.
c. If $\int \max(|g|, |h|, |g - h|)\, dx = 0$, then $g = h$ almost everywhere and $d_N(f, g) = d_N(f, h)$; thus, the triangle inequality still holds.

Next we consider the case where none of the three denominators is zero.

$$D_N(f, g) + D_N(g, h)$$

$$
\begin{aligned}
&= \frac{((\int (f-g)^+\,dx)^p + (\int (f-g)^-\,dx)^p)^{\frac{1}{p}}}{\mathbf{M}(f,g)} + \frac{((\int (g-h)^+\,dx)^p + (\int (g-h)^-\,dx)^p)^{\frac{1}{p}}}{\mathbf{M}(g,h)} \\
&\geq \left(\left(\frac{\int (f-g)^+\,dx}{\mathbf{M}(f,g)} + \frac{\int (g-h)^+\,dx}{\mathbf{M}(g,h)}\right)^p + \left(\frac{\int (f-g)^-\,dx}{\mathbf{M}(f,g)} + \frac{\int (g-h)^-\,dx}{\mathbf{M}(g,h)}\right)^p\right)^{\frac{1}{p}} \\
&\geq \left(\left(\frac{\int (f-g)^+ + (g-h)^+\,dx}{\mathbf{M}^*(f,g,h)}\right)^p + \left(\frac{\int (f-g)^- + (g-h)^-\,dx}{\mathbf{M}^*(f,g,h)}\right)^p\right)^{\frac{1}{p}} \\
&= (I^p + J^p)^{\frac{1}{p}},
\end{aligned}
$$

where

$$
I = \frac{\int (f-g)^+ + (g-h)^+\,dx}{\mathbf{M}^*(f,g,h)} \quad \text{and} \quad J = \frac{\int (f-g)^- + (g-h)^-\,dx}{\mathbf{M}^*(f,g,h)}.
$$

Let $\Gamma(f,g,h) = \int (\max(|g|, |f-g|, |g-h|) - \max(|f|, |h|, |f-h|))^+\,dx$. By subtracting $\Gamma(f,g,h)$ from the numerator and denominator of I at the same time, it follows that

$$
\begin{aligned}
I &\geq \frac{\int (f-g)^+ + (g-h)^+\,dx - \Gamma(f,g,h)}{\mathbf{M}^*(f,g,h) - \Gamma(f,g,h)} \\
&= \frac{\int (f-g)^+ + (g-h)^+\,dx - \Gamma(f,g,h)}{\mathbf{M}(f,h)} \\
&\geq \frac{\int (f-h)^+\,dx}{\mathbf{M}(f,h)}.
\end{aligned}
$$

The first of the above inequalities holds since we are subtracting a non-negative number no greater than the non-negative numerator from the top and bottom while the fraction stays in $[0,1]$. The equality holds due to Lemma 9.4 and the last inequality due to Lemma 9.3. By analogy it can be shown that

$$
J \geq \frac{\int (f-h)^-\,dx}{\mathbf{M}(f,h)}.
$$

Thus, we have $d_N(f,g) + d_N(g,h) \geq d_N(f,h)$.

For general functions $f, g, h \in L(\mathbb{R})$, we can exclude the set $(|f| = \infty) \cup (|g| = \infty) \cup (|h| = \infty)$, since the set where the functions are infinite is of measure zero. Thus, the theorem would proceed the same way since Lemma 9.2 and Lemma 9.3 still hold for $|f|, |g|, |h| < \infty$.

We now show that $d_N \in [0,1]$. Since $d_N(\cdot,\cdot)$ is nonnegative, we only need to show that it is bounded by 1. Applying the Minkowski inequality we have that

$$
\begin{aligned}
&\left((\int (f-g)^+\,dx)^p + (\int (f-g)^-\,dx)^p\right)^{\frac{1}{p}} \\
\leq &\int (f-g)^+\,dx + \int (f-g)^-\,dx \\
= &\int |f-g|\,dx.
\end{aligned}
$$

Thus the numerator is bounded by the denominator and the fraction is no greater than 1. With $\frac{0}{0} := 0$ for $d_N(\cdot,\cdot)$, we have shown that this metric is in $[0,1]$. $\square$

Proof of Proposition 4.1. Since the proposition holds for $p = 1$ and $p = \infty$, we only consider cases when $p > 1$. Consider $\mathbf{x}, \mathbf{y} \in \mathbb{R}^k$. First we show that $d_{\mathrm{M}}(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{y})$. We have that

$$\begin{aligned} d_{\mathrm{M}}(\mathbf{x}, \mathbf{y})^p &= \sum |x_i - y_i|^p \\ &= \sum_{i:x_i \geq y_i} |x_i - y_i|^p + \sum_{i:x_i < y_i} |x_i - y_i|^p \\ &\leq (\sum_{i:x_i \geq y_i} x_i - y_i)^p + (\sum_{i:x_i < y_i} y_i - x_i)^p \\ &= d(\mathbf{x}, \mathbf{y})^p \end{aligned}$$

The inequality holds since the bases of the exponents are positive. Now we show the other inequality holds, that is

$$\begin{aligned} d(\mathbf{x}, \mathbf{y})^p &= (\sum_{i:x_i \geq y_i} x_i - y_i)^p + (\sum_{i:x_i < y_i} y_i - x_i)^p \\ &\leq \{[k^{1/P^*}(\sum_{i:x_i \geq y_i} (x_i - y_i)^p)^{1/p}]^p + [k^{1/P^*}(\sum_{i:x_i < y_i} (y_i - x_i)^p)^{1/p}]^p\} \\ &= k^{p/p^*} \sum |x_i - y_i|^p \\ &= k^{p/p^*} d_{\mathrm{M}}(\mathbf{x}, \mathbf{y})^p, \end{aligned}$$

with $p^* = p/(p-1)$. The inequality holds since $|\sum_{i=1}^m a_i| = |(1, \ldots, 1) \cdot \mathbf{a}| \leq \|(1, \ldots, 1)\|_{p^*} \|\mathbf{a}\|_p$, where $\|\cdot\|_p$ represents the L^p norm, thanks to Hölder's inequality. □

Proof of Theorem 3.7. By definition, a metric space (X, d) is complete if all Cauchy sequences in X converge in X; that is, if the limit point of every Cauchy sequence in X remains in X. Let us first consider a Cauchy sequence in $(L(\mathbb{R}), d)$, where for a given $\epsilon > 0$, there exists some $N > 0$ such that $d(f_n, f_m) < \epsilon$ for all $n, m \geq N$; i.e., $(\int (f_n - f_m)^- \, dx)^p + (\int (f_n - f_m)^+ \, dx)^p < \epsilon^p$. It follows that $\int (f_n - f_m)^- \, dx < \epsilon$ and $\int (f_n - f_m)^+ \, dx < \epsilon$ and thus $\int |f_n - f_m| \, dx < 2\epsilon$. Therefore, $\{f_n\}$ is a Cauchy sequence in L^1 space, where the metric in L^1 is $d(f, g) = \int |f - g| \, dx$ for integrable functions and thus f_n converges to a function f in L^1 by the completeness of L^1 space.

Now we look at a Cauchy sequence $\{f_n\}$ in $(L(\mathbb{R}), d_N)$. By Lemma 9.5 we have that $\int |f_n| \, dx \leq M$ for all n for some positive constant M. It follows that for any given $\epsilon > 0$, there exists some integer $N_0 > 0$ such that $d_N(f_n, f_m) < \epsilon$ for all $n, m \geq N_0$, or in other words, $d(f_n, f_m) < 2M\epsilon$. Therefore, $\{f_n\}$ is Cauchy in $(L(\mathbb{R}), d)$ and by previous results we know that $\{f_n\}$ has a limit in L^1 and therefore $(L(\mathbb{R}), d_N)$ is complete. □

10. Real-valued and text data

Real-valued data sets were downloaded from UCI machine learning repository. The three numbers in the parenthesis after each data set listed below correspond to (number of classes, number of instances, number of features): *airfoil* (2, 1503, 5), *banknote* (2, 1372, 4), *cardiotocography* (10, 2126, 21), *concrete* (2, 1030, 8), *eyestate* (2, 14980, 14), *faults* (7, 1941, 27), *fertility* (2, 100, 9), *gas* (6, 13910, 128), *glass* (6, 214, 10), *housing* (2, 506, 13), *ionosphere* (2, 351, 34), *iris* (3, 150, 4), *landsat* (6, 6435, 36), *leaf* (30, 340, 14), *pageblock* (5, 5473, 10), *pendigits* (10, 10992, 16), *pima* (2, 768, 8), *retinopathy* (2, 1151, 19), *seeds* (3, 210, 7), *segment* (7, 2310, 19), *shuttle* (7, 58000, 9), *sonar* (2, 208, 60), *spambase* (2, 4601,

57), *transfusion* (2, 748, 4), *vertebral* (2, 310, 6), *waveform* (3, 5000, 40), *wdbc* (2, 569, 30), *wilt* (2, 4839, 5), *winequality* (7, 6497, 11), *yeast* (10, 1484, 8). *Concrete* and *housing* were converted to binary classification tasks based on the target mean.

Among the ten text document data sets, *lifesci5* (5, 9000, 52757) was extracted from a collection of abstracts from 5 life sciences journals: *Scientific Reports*, *Oncotarget*, *Proceedings of the National Academy of Sciences of the United States of America*, *ACS Applied Materials and Interfaces* and *PloS One*, obtained from the Europe PMC life science database. As for *reuters* (4, 2065, 8943), documents were collected by selecting the following four topics exclusively: "interest", "trade", "grain" and "crude". Each of these data sets was pre-processed by stop-word removal followed by Porter stemming. The remaining pre-processed data sets were *webkb* (4, 2803, 7288) downloaded from (Cardoso-Cachopo, 2007), *20newsgroups* (20, 4000, 130107) from scikit-learn python library, *moviereview* (2, 2000, 39659) from (Pang and Lee, 2004), *farm-ads* (2, 4143, 54877), *NIPS* (4, 5811, 11463), *TTC3600* (6, 3600, 5692) from the UCI Machine Learning Repository, *sdm06* (27, 930, 99899) from (Dalkilic et al., 2006) and *bbc* (5, 2225, 9635) from (Greene and Cunningham, 2006). Note that for the NIPS data set, we made 4 classes of papers according to their publication year: 1987-1994, 1995-2001, 2002-2008, 2009-2015. All text document data sets used tf-idf as features.

11. Protein function data

Protein function data were downloaded from the UniProt database (July 2015). In particular, we collected protein functions for the following model organisms where sufficient annotations are available: *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Escherichia coli*. Only those annotations with (experimental) evidence codes EXP, IDA, IMP, IPI, IGI, IEP, TAS, and IC were considered. Table 1 summarizes the data sets: here the genome size corresponds to the total number of proteins available for each species in UniProt. The following three columns show the number of proteins that are annotated in the three domains of Gene Ontology accordingly.

Organism	Genome size	MFO	BPO	CCO
H. sapiens	20,193	11,979	11,398	12,691
M. musculus	16,733	6,728	7,702	7,322
A. thaliana	14,305	4,266	5,749	5,950
S. cerevisiae	6,720	4,051	4,676	4,102
E. coli	4,433	2,272	2,331	2,119

Table 1. Data set sizes for the five organisms used in this work. The genome size refers to the number of protein sequences available in UniProt for each species. Other numbers refer to the proteins with experimental annotations in each of the three concept hierarchies.

The conditional probability tables were estimated using the maximum likelihood approach from the entire set of functionally annotated proteins in UniProt. This set included 72977 proteins from 1576 species with MFO terms, 92874 proteins from 1503 species with BPO terms, and 89693 proteins from 862 species with CCO terms.

12. Calculation of protein sequence similarity

The sequence similarity of two protein sequences was measured as the ratio of the number of identical characters in the alignment and the length of the longer protein sequence. Sequence alignment was obtained using a Needleman-Wunsch algorithm with BLOSUM62 similarity matrix, gap opening penalty of 11 and gap extension penalty of 1.

13. Phylogenetic clustering

The functional phylogenetic trees with respect to a group of organisms are generated using single-linkage hierarchical clustering (Tan et al., 2006). This algorithm starts by considering every data point (species) to be a cluster of one element and in each step merges the two closest clusters. The algorithm continues until all original data points belong to the same cluster. The distance between species is based on pairwise distances between functionally annotated proteins as described below. For simplicity, we use normalized semantic distance from Eq. 10 with $p = 1$ in all experiments.

Without loss of generality, we illustrate the species distance calculation by showing how to compute the distances between *A. thaliana* (A) and all other organisms using protein function data only. An important challenge in this task arises from unequal genome sizes as well as unequal fractions of experimentally annotated proteins in each species (Table 1), making most distance calculation techniques unsuitable for this task. We therefore carry out sampling to compare species using a fixed yet sufficiently large set of N proteins from each species. The algorithm first samples (with replacement) $N = 1000$ proteins from each species. It then counts the number of times the proteins from *E. coli* (E), *H. sapiens* (H), *M. musculus* (M) and *S. cerevisiae* (Y) are functionally most similar to proteins in *A. thaliana*, with ties resolved uniformly randomly. These counts are used to calculate the directional distances between *A. thaliana* and the remaining four species. The procedure is repeated $B = 1000$ times with different bootstrap samples to stabilize the results. The details of the algorithm are shown in Algorithm 1.

We experimented with clustering using single linkage, complete linkage, and group-average strategies for computing distances between clusters. We noticed little change in the resulting functional phylogenies for either ontology. There was also no dependence on the selection of N in Algorithm 1, where we evaluated $N = 500$, $N = 1000$, and $N = 1500$. Note that N was required to be smaller than the smallest set in Table 1, which was *E. coli* in each ontology.

14. Additional results

We carried out additional experiments in order to further evaluate proposed distance functions. These experiments investigated the influence of data normalization and performance assessment criteria. Figure 10-14 show the compairison against both metrics and non-metrics. The normalization techniques we use include: (i) *z-score*, i.e., standardization per feature (ii) *min-max*, i.e., re-scaling each feature to be within $[0, 1]$ (iii) *unit*, i.e., re-scaling each data point to the unit length.

Figures 10-13 show the number of bootstrapped wins for all dissimilarity functions over 30 real-valued data sets under various normalizations. The top-performing distance measure on these relatively low-dimensional problems was the cityblock distance, which belongs to both L^p and d^p family when $p = 1$, followed by d^2 and d^2_N metrics. Importantly, we also find

Algorithm 1: (from Section 5.2) Computing distances from *A. thaliana* (A) to *E. coli* (E), *H. sapiens* (H), *M. musculus* (M), and *S. cerevisiae* (Y) respectively.

Input : Sets of protein functions $X = \{\text{Prot}_k^X, k = 1, \ldots, n_X\}$, where n_X is the number of functionally annotated proteins in organism X, for $X \in \mathcal{S} = \{A, E, H, M, Y\}$ and a metric d_N on ontologies.

Output: Distances d_{AE}, d_{AH}, d_{AM} and d_{AY}.

begin

  Initialize the bootstrapping sample size $N > 0$ and iteration counts $B > 0$.

  for $b = 1, 2, \ldots B$ **do**

    Bootstrap N proteins Prot_i^X, $i = 1, 2, \ldots, N$ from each organism $X \in \mathcal{S}$.;

    for $i = 1, 2, \ldots, N$ **do**

      for $X \in \mathcal{S} - A$ **do**

        $d(\text{Prot}_i^A, X) := \min_{j \in \{1,\ldots,N\}} d_N(\text{Prot}_i^A, \text{Prot}_j^X)$;

      end

      $I(i) = \underset{X \in \mathcal{S}-A}{\text{argmin}}\, d(\text{Prot}_i^A, X)$

    end

    for $X \in \mathcal{S} - A$ **do**

      $d_X(b) = 1 - \frac{1}{N} \cdot |\{I == X\}|$;

    end

  end

  for $X \in \mathcal{S} - A$ **do**

    $d_{AX} = \frac{1}{B} \sum_{b=1}^{B} d_X(b)$;

  end

end

that the d^p and d_N^p metrics outperform their L^p counterparts when $p > 1$. Figure 14 shows another tournament plot for text document data sets under the unit normalization. Similar to Figure 2 in the main text, we see that d_N^1 and d_N^2 are among the best performing functions over all metric and non-metric dissimilarities. The normalized Euclidean distance performed well among metrics and Spearman's distance and normalized cityblock performed well among non-metrics. These results are provided as a spreadsheet in Electronic Supplementary Materials. Although the effect of data normalization deserves attention, all conclusions reached in the main portion of the paper remain unchanged.

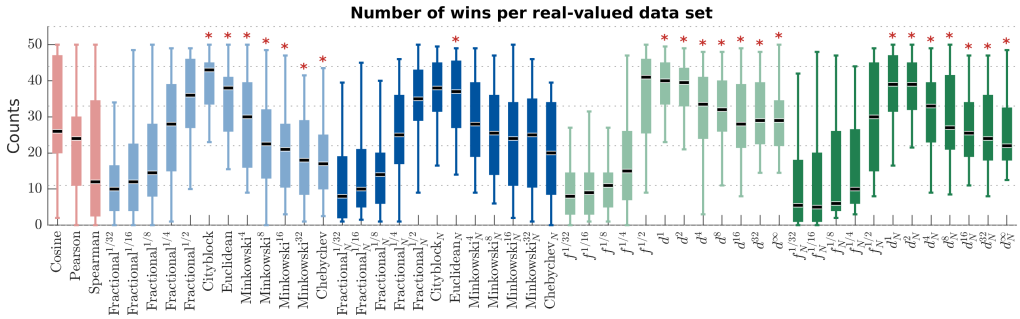

FIGURE 10. Comparison between dissimilarities for real-valued data using z-score normalization. The functions are color coded as follows: L^p family (light blue), normalized L^p family (dark blue), d^p family (light green), normalized d^p family (dark green), cosine distance and the correlation coefficients (light red). All metrics are labeled by an asterisk.

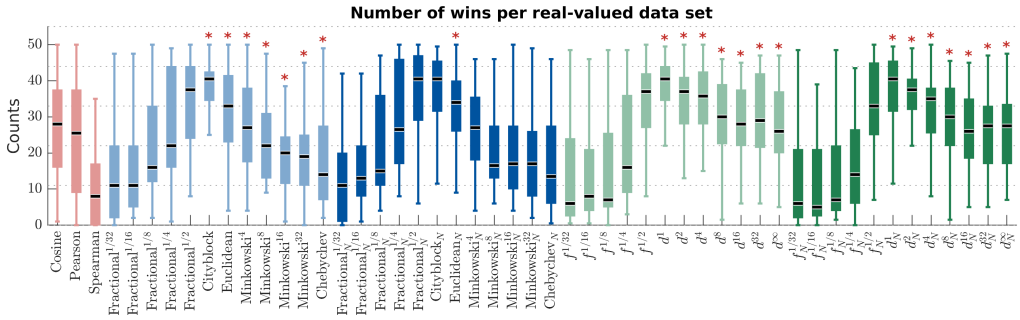

FIGURE 11. Comparison between dissimilarities for real-valued data using min-max normalization. The functions are color coded as follows: L^p family (light blue), normalized L^p family (dark blue), d^p family (light green), normalized d^p family (dark green), cosine distance and the correlation coefficients (light red). All metrics are labeled by an asterisk.

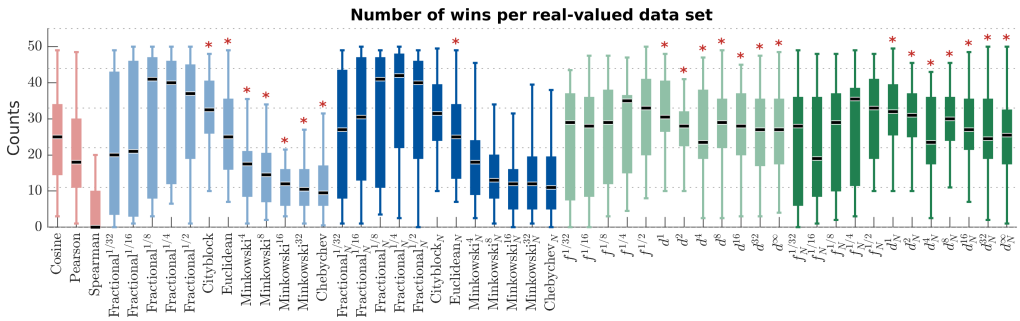

FIGURE 12. Comparison between dissimilarities for real-valued data using unit normalization. The functions are color coded as follows: L^p family (light blue), normalized L^p family (dark blue), d^p family (light green), normalized d^p family (dark green), cosine distance and the correlation coefficients (light red). All metrics are labeled by an asterisk.

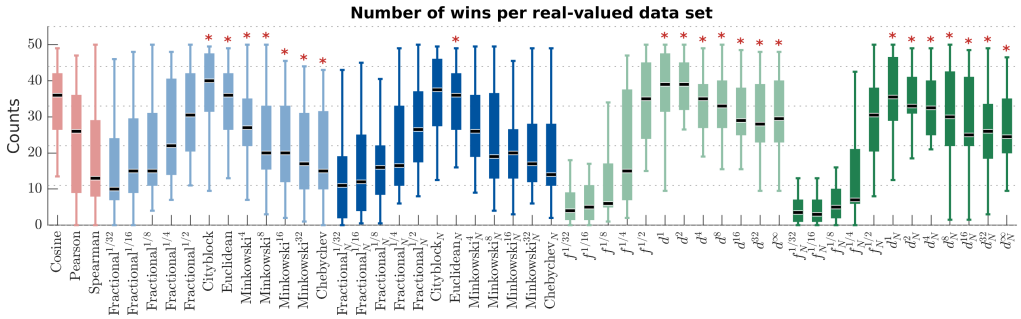

FIGURE 13. Comparison between dissimilarities for real-valued data using z-score normalization and followed by unit normalization. The functions are color coded as follows: L^p family (light blue), normalized L^p family (dark blue), d^p family (light green), normalized d^p family (dark green), cosine distance and the correlation coefficients (light red). All metrics are labeled by an asterisk.

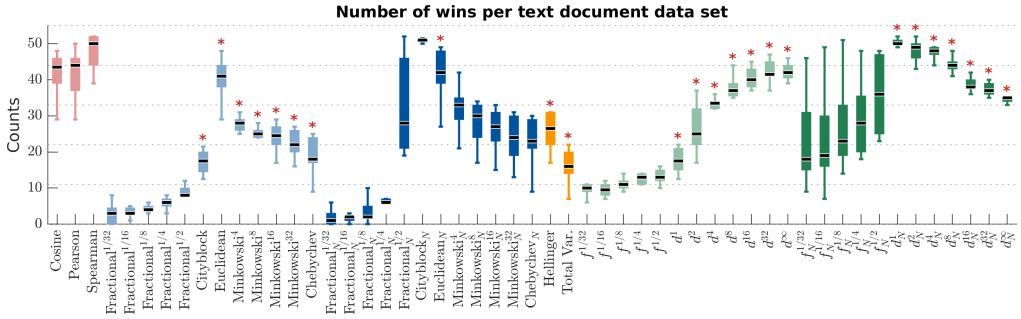

FIGURE 14. Comparison between dissimilarities for text data using unit normalization. The functions are color coded as follows: L^p family (light blue), normalized L^p family (dark blue), d^p family (light green), normalized d^p family (dark green), cosine distance and the correlation coefficients (light red). All metrics are labeled by an asterisk.