The Similarity Metric and the Distance Metric

Bin Ma Kaizhong Zhang
Department of Computer Science
University of Western Ontario
London, Ont. N6A 5B7
Canada
{bma,kzhang}@csd.uwo.ca

Abstract

Distance metric is important in many 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. We propose a similarity metric definition and show its relationship with distance metric. We also present general solutions of normalizing distance metric and similarity metric.

Introduction

Distance and similarity measures are widely used in bioinformatics research and other fields. For examples sequence edit distance and tree edit distance are used in many areas [4, 11], distance metrics are used in constructing phylogenetic trees, distance metrics are used for improving database search [6], and protein sequence similarity based on blosum matrix is used for protein sequence comparison [7].

Distance metric is a well defined concepts. In contrast, although similarity measures are widely used and its properties are studied and discussed [9, 8], it seems that there is no formal definition for the concept. In this paper, we propose a formal definition of similarity metric and show its properties and its relation to distance metric.

We also consider the problem of normalized distance metric and normalized similarity metric. Although there are studies on normalize specific distance metrics [1, 2], there is no general solution. We give general formulas to normalized a similarity metric or a distance metric.

The Similarity metric and the distance metric

Recall that a distance metric on a set X is a non-negative function $d(x, y)$ on the Cartesian product $X \times X$ of X satisfying the following properties for any $x, y, z \in X$:

1. $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$

We say a distance metric $d(x, y)$ is a normalized distance metric if $d(x, y) \leq 1$.

The Similarity metric definition

Given a set X, a function $s(x, y)$ which assigns a real value to every pair $x, y \in X$ is a similarity metric if, for any $x, y, z \in X$, it satisfies the following properties:

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

The first property says that $s(x, y)$ is symmetric. The second property says that for any x the self similarity is non-negative. The third property says that for any x the self similarity is greater than or equals to the similarity between x and any y. The fourth property states that the similarity between x and z through y is less or equals to the direct similarity between x and z plus self similarity of y. This property is the equivalent of triangle inequality in distance metric. The last proper states that if $s(x, x) = s(y, y) = s(x, y)$ then $x = y$ which is justified by the following lemma.

Lemma 1. *Let $s(x, y)$ be a real function satisfying similarity metric condition 1 through 4, if $s(x, x) = s(y, y) = s(x, y)$, then for any z $s(x, z) = s(y, z)$.*

Proof. From $s(x,y) + s(y,z) \leq s(y,y) + s(x,z)$, we have $s(y,z) \leq s(x,z)$ and from $s(y,x) + s(x,z) \leq s(x,x) + s(y,z)$, we have $s(x,z) \leq s(y,z)$. □

We say a similarity metric $s(x,y)$ is a normalized similarity metric if $|s(A,B)| \leq 1$.

Relationship between the similarity metric and the distance metric

We now consider the relationship between the similarity metric and the distance metric.

Lemma 2. *Given a similarity metric $s(x,y)$, $d(x,y) = \frac{s(x,x)+s(y,y)}{2} - s(x,y)$ is a distance metric.*

Proof. Since $s(x,x) \geq s(x,y)$ and $s(y,y) \geq s(x,y)$, $d(x,y) \geq 0$. If $d(x,y) = 0$, then $s(x,x) = s(y,y) = s(x,y)$ which implies that $x = y$. Since $s(x,y)$ is symmetric, then $d(x,y)$ is also symmetric. For triangle inequality, we have

$$\begin{aligned} & d(x,z) \\ = & \frac{s(x,x)+s(z,z)}{2} - s(x,z) \\ = & \frac{s(x,x)+s(z,z)}{2} + s(y,y) - s(x,z) - s(y,y) \\ \leq & \frac{s(x,x)+s(z,z)}{2} + s(y,y) - s(x,y) - s(y,z) \\ = & d(x,y) + d(y,z) \end{aligned}$$

Therefore $d(x,y)$ is a distance metric. □

Lemma 3. *Given a distance metric $d(x,y)$, for any $k \geq 1$, $s_k(x,y) = \frac{d(x,o)+s(y,o)}{k} - d(x,y)$, where o is an arbitrary element, is a similarity metric.*

Proof. It is clear that $s_k(x,y) = s_k(y,x)$ and $s_k(x,x) \geq 0$. Since $k \geq 1$, with triangle inequality, it is easy to show $s_k(x,y) \leq s_k(x,x)$. From $s_k(x,x) = s_k(y,y) = s_k(x,y)$, we have $d(x,y) = 0$ which means that $x = y$. Finally, we have

$$\begin{aligned} & s_k(x,y) + s_k(y,z) \\ = & \frac{d(x,o)+d(y,o)}{k} - d(x,y) + \frac{d(y,o)+d(z,o)}{k} - d(y,z) \\ = & \frac{d(x,o)+d(z,o)}{k} + \frac{d(y,o)+d(y,o)}{k} - (d(x,y) + d(y,z)) \\ \leq & \frac{d(x,o)+d(z,o)}{k} + \frac{d(y,o)+d(y,o)}{k} - (d(x,z) + d(y,y)) \\ = & \frac{d(x,o)+d(z,o)}{k} - d(x,z) + \frac{d(y,o)+d(y,o)}{k} - d(y,y) \\ = & s_k(x,z) + s_k(y,y) \end{aligned}$$

Therefore $s_k(x,y)$ is a similarity metric. □

The relationship between a normalized similarity metric and a distance metric is as follows.

Lemma 4. *If $s(x,y)$ is a normalized similarity metric, then $\frac{1}{2}(1-s(x,y))$ is a normalized distance metric. If $s(x,y) \geq 0$ is a normalized similarity metric, then $(1-s(x,y))$ is a normalized distance metric.*

Lemma 5. *If $d(x,y)$ is a normalized distance metric, then $1-d(x,y)$ is a normalized similarity metric.*

The normalized similarity metric and distance metric

We first present results concerning normalized similarity metric We then use those results and the relationship between distance metric and similarity metric from the previous section to show the results concerning normalized distance metric.

The normalized similarity metric

Given a similarity metric $s(x,y)$, we now show how to generate a normalized similarity metric base on $s(x,y)$.

Theorem 1. *Suppose that $s(x,y)$ is a non negative similarity metric, then*

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

is a normalized similarity metric.

Proof. It is clear that $S(x,y) = S(x,y)$, $S(x,x) \geq 0$, and $S(x,x) \geq S(x,y)$. Also since $S(x,x) = S(y,y) = S(x,y)$ implies $s(x,x) = s(y,y) = s(x,y)$, we know that $x = y$.

We now show that $S(x,y) + S(y,z) \leq S(y,y) + S(x,z)$. There are three cases.

1. $s(y,y) \geq s(x,x) \geq s(z,z)$.

$$\begin{aligned} & S(x,y) + S(y,z) \\ = & \frac{s(x,y)}{s(y,y)} + \frac{s(y,z)}{s(y,y)} \\ \leq & \frac{s(x,z)}{s(y,y)} + \frac{s(y,y)}{s(y,y)} \\ \leq & \frac{s(x,z)}{s(x,x)} + \frac{s(y,y)}{s(y,y)} \\ = & S(x,z) + S(y,y) \end{aligned}$$

2. $s(x,x) \geq s(y,y) \geq s(z,z)$.

$$\begin{aligned}
& S(x,y) + S(y,z) \\
= & \frac{s(x,y)}{s(x,x)} + \frac{s(y,z)}{s(y,y)} \\
= & \frac{s(y,y)s(x,y) + s(x,x)s(y,z)}{s(x,x)s(y,y)} \\
= & \frac{s(y,y)(s(x,y) + s(y,z))}{s(x,x)s(y,y)} \\
& + \frac{(s(x,x) - s(y,y))s(y,z)}{s(x,x)s(y,y)} \\
\leq & \frac{s(y,y)(s(x,z) + s(y,y))}{s(x,x)s(y,y)} \\
& + \frac{(s(x,x) - s(y,y))s(y,y)}{s(x,x)s(y,y)} \\
= & \frac{s(y,y)s(x,z) + s(x,x)s(y,y)}{s(x,x)s(y,y)} \\
= & \frac{s(x,z)}{s(x,x)} + \frac{s(y,y)}{s(y,y)} \\
= & S(x,z) + S(y,y)
\end{aligned}$$

3. $s(x,x) \geq s(z,z) \geq s(y,y)$.

$$\begin{aligned}
& S(x,y) + S(y,z) \\
= & \frac{s(x,y)}{s(x,x)} + \frac{s(y,z)}{s(z,z)} \\
= & \frac{s(z,z)s(x,y) + s(x,x)s(y,z)}{s(x,x)s(z,z)} \\
= & \frac{s(z,z)(s(x,y) + s(y,z))}{s(x,x)s(z,z)} \\
& + \frac{(s(x,x) - s(z,z))s(y,z)}{s(x,x)s(z,z)} \\
\leq & \frac{s(z,z)(s(x,z) + s(y,y))}{s(x,x)s(z,z)} \\
& + \frac{(s(x,x) - s(z,z))s(y,y)}{s(x,x)s(z,z)} \\
= & \frac{s(z,z)s(x,z) + s(x,x)s(y,y)}{s(x,x)s(z,z)} \\
= & \frac{s(x,z)}{s(x,x)} + \frac{s(y,y)}{s(z,z)} \leq \frac{s(x,z)}{s(x,x)} + \frac{s(y,y)}{s(y,y)} \\
= & S(x,z) + S(y,y)
\end{aligned}$$

□

Theorem 2. *Suppose that $s(x,y)$ is a similarity metric, then*

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

is a normalized similarity metric.

Proof. We omit the proof due to the page limit. □

The normalized distance metric

Theorem 3. *Suppose that $d(x,y)$ is a distance metric, then for any element o*

$$\frac{d(x,y)}{2\max\{d(x,o), d(y,o)\}} - \frac{\min\{d(x,o), d(y,o)\}}{2\max\{d(x,o), d(y,o)\}} + \frac{1}{2}$$

is a normalized distance metric.

Proof. Let $d(x,y)$ be a distance metric and $s(x,y) = d(x,o) + d(y,o) - d(x,y)$, then $s(x,y)$ is a similarity metric and $s(x,y) \geq 0$ by triangle inequality.

Therefore $S(x,y) = \frac{s(x,y)}{\max\{s(x,x), s(y,y)\}}$ is a normalized similarity metric and $1 - S(x,y)$ is a normalized distance metric.

Therefore

$$\begin{aligned}
& 1 - \frac{s(x,y)}{\max\{s(x,x), s(y,y)\}} \\
= & \frac{2\max\{d(x,o), d(y,o)\} - d(x,o) - d(y,o)}{2\max\{d(x,o), d(y,o)\}} \\
& + \frac{d(x,y)}{2\max\{d(x,o), d(y,o)\}} \\
= & \frac{\max\{d(x,o), d(y,o)\} - \min\{d(x,o), d(y,o)\}}{2\max\{d(x,o), d(y,o)\}} \\
& + \frac{d(x,y)}{2\max\{d(x,o), d(y,o)\}} \\
= & \frac{d(x,y)}{2\max\{d(x,o), d(y,o)\}} \\
& - \frac{\min\{d(x,o), d(y,o)\}}{2\max\{d(x,o), d(y,o)\}} + \frac{1}{2}
\end{aligned}$$

is a normalized distance metric. □

Theorem 4. *Suppose that $d(A,B)$ is a distance metric, then for any element o and $k \geq 1$*

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

is a normalized distance metric.

Proof. Let $d(x,y)$ be a distance metric and $s(x,y) = \frac{d(x,o)+d(y,o)}{k} - d(x,y)$, then $s(x,y)$ is a similarity metric.

Therefore $S(x,y) = \frac{s(x,y)}{s(x,x)+s(y,y)-s(x,y)}$ is a normalized similarity metric and $\frac{1}{2}(1 - S(x,y))$ is a normalized distance metric.

Therefore

$$\begin{aligned}
& \frac{1}{2}(1 - \frac{s(x,y)}{s(x,x) + s(y,y) - s(x,y)}) \\
= & \frac{d(x,y)}{d(x,y) + \frac{d(x,O)+d(y,O)}{k}}
\end{aligned}$$

is a normalized distance metric. □

Applications

Set similarity and distance metric

It is easy to show that $|A \cap B|$ is a similarity metric and therefore, from the results of previous sections, we have the following results.

- $|A \cap B|$ is a similarity metric.
- $|A \cup B| - |A \cap B|$ is a distance metric.
- $\frac{|A \cap B|}{\max\{|A|,|B|\}}$ is normalized similarity metric.
- $\frac{\max\{|A-B|,|B-A|\}}{\max\{|A|,|B|\}}$ is normalized distance metric.
- $\frac{|A \cap B|}{|A \cup B|}$ is normalized similarity metric.
- $\frac{|A-B|+|B-A|}{|A \cup B|}$ is normalized distance metric.

Information theoretic similarity and distance metric

Let $H(X)$ be the entropy of random variable X, $H(X|Y)$ be the conditional entropy, and $M(X,Y)$ be the mutual information between variables X and Y, then it is easy to see that $M(x,y)$ is a similarity metric and from the results of previous sections, we have the following results.

- $M(X,Y)$ is a similarity metric.
- $H(X|Y) + H(Y|X)$ is a distance metric.
- $\frac{M(X,Y)}{\max\{H(X),H(Y)\}}$ is a normalized similarity metric.
- $\frac{\max\{H(X|Y),H(Y|X)\}}{\max\{H(X),H(Y)\}}$ is a normalized distance metric.
- $\frac{M(X,Y)}{H(X,Y)}$ is normalized similarity metric.
- $\frac{H(X|Y)+H(Y|X)}{H(X,Y)}$ is a normalized distance metric.

Sequence edit distance

It is well know that if the cost for basic operations of insertion, deletion, and substitution, is a distance metric, then the sequence edit distance $d(s,t)$, defined between two sequences s and t derived from finding the minimum cost operation sequence that transforms s to t, is also a distance metric.

Several normalized edit distances have been proposed and studied [3, 5]. Examples are $\frac{d(s,t)}{|s|+|t|}$, $\frac{d(s,t)}{\max\{|s|,|t|\}}$, and $n(s,t) = \min\{\frac{p(s,t)}{|p|} | p \text{ is a path that change } s \text{ to } t\}$. Although these are referred to as normalized edit distance, they are not distance metric.

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

$$\frac{d(s,t)}{2\max\{|s|,|t|\}} - \frac{\min\{|s|,|t|\}}{2\max\{|s|,|t|\}} + \frac{1}{2}$$

Acknowledgements

We thank the support from NSERC grants and a sharcnet fellowship.

References

[1] H. Bunke and K. Shearer, 'A graph distance metric based on the maximal common subgraph', *Pattern. Recogn. Lett.*, 19 (3-4), pp. 255-259, 1998.

[2] M. Li, X. Chen, X. Li, B. Ma, P. Vitanyi, 'The similarity metric', *Proc. of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 863-872, 2003.

[3] A. Marzal and E. Vidal, 'Computation of normalized distance and applications', *IEEE trans. on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 926-932, 1993.

[4] S.E. Needleman and C.D. Wunsch, 'A general method applicable to the search for similarities in the amino-acid sequences of two proteins', *J. Mol. Bio.*, 48, pp.443-453, 1970.

[5] B.J. Oommen and K. Zhang, 'The normalized string editing problem revisited', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 669-672, 1996.

[6] C. Sahinalp, M. Tasan, J. Macker, and M. Ozsoyoglu, 'Distance Based Indexing for String Proximity Search', *19th International Conference on Data Engineering*, 2003.

[7] T.F. Smith and M.S. Waterman, 'Comparison of biosequences', *Adv. in Appl. Math.* 2, pp.482-489, 1981

[8] A. Stojmirovic and V. Pestov, 'Indexing schemes for similarity search in datasets of short protein fragments', *ArXiv e-print cs.DS/0309005*, 2003.

[9] M.S. Waterman, T.F. Smith, and W.A. Beyer, 'Some biological sequence metrics', *Adv. in Math.* 20, pp.367-387, 1976

[10] K. Zhang, 'Computing similarity between RNA secondary structures', *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, pp. 126-132, 1998.

[11] K. Zhang and D. Shasha, 'Simple fast algorithms for the editing distance between trees and related problems', *SIAM J. Computing* vol. 18, no. 6, pp.1245-1262, 1989