Discrete Applied Mathematics 167 (2014) 310–314

Contents lists available at ScienceDirect

Discrete Applied Mathematics

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

Note

Distances between sets based on set commonality

K.J. Horadam [*], M.A. Nyblom

SMGS, RMIT University, Melbourne, Australia

ARTICLE INFO

Article history:
Received 20 August 2012
Received in revised form 22 August 2013
Accepted 30 October 2013
Available online 28 November 2013

Keywords:
Metric
Distance function
Distance between sets
Set intersection
Jaccard distance
Minkowski distance
Graph matching
Pattern matching
Biometric verification

ABSTRACT

We construct a new family of normalised metrics for measuring the dissimilarity of finite sets in terms of the sizes of the sets and of their intersection. The family normalises a set-based analogue of the Minkowski metric family. It is parametrised by a real variable $p \geq 1$, is monotonic decreasing in p, equals the normalised set difference metric when $p = 1$ and equals the normalised maximum difference metric in the limit $p \to \infty$. These metrics are suitable for comparison of finite sets in any context. Several applications to comparison of finite graphs are described.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In many pattern matching, data mining, coding, biometric and more general artificial intelligence and image processing applications one of the underlying ideas is to determine the closeness or commonality between a pair of objects and to find the nearest neighbour or best match to an object (feature or datavector) of interest, in a collection of such objects. The definition of nearness depends on the object type, and there are many different distance functions for different types of objects, some of which are metrics. For example, the Taxicab, Euclidean, Minkowski and Chebychev metrics are used for numerical data, the Hamming metric for binary or categorical data, the edit distance for strings and the graph edit distance for graphs. The modified Hausdorff distance is used for point set matching and the Jaccard metric for sample set comparisons. For feature matching, a similarity coefficient is commonly used instead. An encyclopedic listing of known distances and similarities appears in [2].

Here we are interested in distances, like the set difference metric, which describe the dissimilarity between two finite sets (which may consist of any type of elements) in terms purely of the sizes of the sets and of their intersection. The existence of such distances is of intrinsic mathematical interest but has applications in many of the above-mentioned areas.

Our interest is motivated by the problem of comparing pairs of sets of subgraphs in a pair of graphs. This situation is more general than either comparing pairs of point sets or comparing pairs of graphs.

Our main result (Theorem 8) is the construction of a family of normalised metrics for measuring the dissimilarity of finite sets in terms of the sizes of the sets and of their intersection. It is parametrised by a real variable $p \geq 1$, is monotonic decreasing in p, equals the normalised set difference (Jaccard) metric when $p = 1$ and equals the normalised maximum difference metric in the limit $p \to \infty$. In the process we also find normalised distances which are not metrics but which have advantages in some applications.

[*] Corresponding author. Tel.: +61 3 9925 2283; fax: +61 3 9925 2454.
E-mail address: kathy.horadam@rmit.edu.au (K.J. Horadam).

0166-218X/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.dam.2013.10.037

The remainder of the paper is organised as follows. We complete this section with known examples of distances between sets which use their sizes and the sizes of their intersections. Section 2 introduces a new normalised distance (which is not a metric) based on the geometric mean of the set sizes, and a new family of metrics, which is a set-based analogue of the Minkowski metric family. Its monotonic behaviour and parametric transformation from the set difference metric to the maximum difference metric are described.

The most important result, Theorem 8, is proved in Section 3. It shows that a normalisation, based on set intersection, of the family found in Section 2 produces a family of metrics. A different normalisation produces a family of distances which are not metrics. These normalised metrics and distances are suitable for the comparison of finite sets in any context. In the final Section 4, several applications in the context of graph matching are discussed.

A *distance* or *dissimilarity* on a data set X is a non-negative, symmetric and reflexive function $d : X \times X \to \mathbb{R}$. It is *normalised* if its values are all ≤ 1. It is a *metric* if for all $u, v, w \in X$ it also satisfies the identity of indiscernibles: $d(u, v) = 0 \Leftrightarrow u = v$; and the *triangle inequality*: $d(u, v) \leq d(u, w) + d(w, v)$.

A small number of known metrics on sets are based purely on the sizes of sets and of their intersections. We use the following notation throughout.

Notation. Let U be a nonempty set. The set X on which distances are defined is the set of all *finite nonempty* subsets of U. This is a subset of the power set $\mathcal{P}(U)$ (set of all subsets) of U. Suppose that $X_i, X_j, X_k \in X$.
Let $x_i = |X_i|$, $x_{ij} = |X_i \cap X_j|$, $x_{ijk} = |X_i \cap X_j \cap X_k|$ etc., and define $m_{ij} = x_i - x_{ij}$, so $x_i > 0$ and $m_{ij} \geq 0$. Further, write $x_i = x_i^* + y_{ij} + y_{ik} + x_{ijk}$, where $y_{ij} = y_{ji} = x_{ij} - x_{ijk}$, so $m_{ij} = x_i - x_{ij} = x_i^* + y_{ik}$.

Remark. Any distance d in what follows can be extended to a distance on $X \cup \{\emptyset\}$ by defining $d(\emptyset, \emptyset) := 0$ and noting that $|\emptyset| = 0$. We restrict to X for simplicity.

The following examples are metrics on X.

Example 1. 1. The set difference metric d_{sd} on X:

$$d_{\text{sd}}(X_i, X_j) = |X_i \cup X_j| - |X_i \cap X_j| = x_i + x_j - 2x_{ij} = m_{ij} + m_{ji}.$$

Division by the size of the set union gives the Jaccard metric on X:

$$\overline{d}_{\text{sd}}(X_i, X_j) = (m_{ij} + m_{ji})/(x_i + x_j - x_{ij}) = 1 - x_{ij}/(x_i + x_j - x_{ij}). \tag{1}$$

2. The maximum difference metric $d_{\max}$ on X:

$$d_{\max}(X_i, X_j) = \max\{|X_i|, |X_j|\} - |X_i \cap X_j| = \max\{m_{ij}, m_{ji}\}.$$

Division by the size of the maximum set gives a normalised metric on X:

$$\overline{d}_{\max}(X_i, X_j) = \max\{m_{ij}, m_{ji}\}/ \max\{x_i, x_j\} = 1 - x_{ij}/ \max\{x_i, x_j\}. \tag{2}$$

3. For each $p \geq 1$, the weighted mean of the above two metrics on X:

$$d_{1,p} = (1/p)d_{\text{sd}} + (1 - 1/p)d_{\max}.$$

The weighted mean of their normalisations is a normalised metric on X:

$$\widehat{d}_{1,p} = (1/p)\overline{d}_{\text{sd}} + (1 - 1/p)\overline{d}_{\max}. \tag{3}$$

The metrics $d_{\max}$ and $\overline{d}_{\max}$ are used in graph matching problems, where the sets are the node sets of the graphs being compared, and they are called the *Zelinka* or *common subgraph* metric and the *Bunke–Shearer* metric, respectively [2].

Given a metric on X, a simple normalisation taking set intersection size into account yields a normalised distance which also satisfies the identity of indiscernibles.

Lemma 2. *If d is a metric on X then $\overline{d}(X_i, X_j) = d(X_i, X_j)/[x_{ij} + d(X_i, X_j)]$ defines a normalised distance on X which satisfies the identity of indiscernibles.* □

It can be shown that for $d_{1,p}$ as in Example 1.3, the normalisation $\overline{d}_{1,p}$ of Lemma 2 is a metric for every $p \geq 1$.

2. New metrics based on set commonality

Other normalisation functions dependent on the sizes of X_i and X_j could be chosen, for instance $\min\{x_i, x_j\}$, or the harmonic or geometric means of x_i and x_j. Ortega et al. [7] found that, for comparing two noisy spatial point patterns X_i and X_j in retinal images for biometric verification, normalisation of the number x_{ij} of matched points by geometric mean $(x_{ij}/\sqrt{x_i x_j})$ is superior to $x_{ij}/ \min\{x_i, x_j\}$. Jeffers et al. [5] found that $x_{ij}/\sqrt{x_i x_j}$ is superior to $x_{ij}/(x_i + x_j - x_{ij})$ for similarity comparison of the same type of data.

This prompted us to derive a distance function for sets from the geometric mean of the set sizes. A check of [2] leads us to believe that it is new.

Download English Version:

https://daneshyari.com/en/article/418791

Download Persian Version:

https://daneshyari.com/article/418791

Daneshyari.com