Generalizing Hamming Distance to Finite Sets

to the purpose of classifying heterogeneous objects

Marc Bezem
Utrecht University
Department of Philosophy *

Maarten Keijzer
Cap Volmac
Adaptive Systems B.V. †

Abstract

We propose a distance measure to compare objects that have heterogeneous sets of characteristics, such as encountered in, for example, medical diagnosis and genetic programming.

1 Introduction

Consider two bitstrings $\vec{s}_1$ and $\vec{s}_2$ of equal length, say n. The *Hamming distance* $d(\vec{s}_1, \vec{s}_2)$ between $\vec{s}_1$ and $\vec{s}_2$ is by definition the number of positions in which $\vec{s}_1$ and $\vec{s}_2$ differ. It is well-known (and, moreover, easily verified) that $d(\vec{s}_1, \vec{s}_2)$ satisfies the following general conditions axiomatising a real-valued metric on a set S, in the special case here the set $\{0,1\}^n$ of bitstrings of length n:

(i) $\forall x, y \in S \;\; (d(x,y) = 0 \iff x = y)$

(ii) $\forall x, y \in S \;\; d(x,y) = d(y,x)$ (symmetry)

(iii) $\forall x, y, z \in S \;\; d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality)

The notion of Hamming distance suggests a scheme for classification in the following straightforward way. Given bitstrings $\vec{s}_1, \ldots, \vec{s}_k$, the set $\{0,1\}^n$ can be partitioned into subsets $N_1, \ldots, N_k$ (so-called neighbourhoods), in such a way that each N_i consists of all bitstrings whose Hamming distance to $\vec{s}_i$ is minimal, as compared with their distance to $\vec{s}_1, \ldots, \vec{s}_k$. (The fact that there may be more than one bitstring among $\vec{s}_1, \ldots, \vec{s}_k$ at minimal Hamming distance, is considered irrelevant for this paper. In such a case, any choice is good enough for our purposes.)

*P.O. Box 80126, 3508 TC Utrecht, The Netherlands, e-mail `bezem@phil.ruu.nl`.
†P.O. Box 2575, 3500 GN Utrecht, The Netherlands, e-mail `mkeijzer@knoware.nl`.

As running example of a classification problem we take medical diagnosis, 'the recognition of a disease from its symptoms' according to [6]. In this context, the clinical picture of a patient should be compared with a number of known syndromes. A syndrome is 'a group of disease symptoms commonly found in association with one another' [6]. The diagnosis is the syndrome that matches best with the clinical picture of the patient.

Bitstrings and Hamming distance provide a rudimentary formalism for diagnosis. Assume we have n boolean valued symptoms, numbered 1 through n. Every syndrome as well as any clinical picture corresponds to a bitstring encoding the presence/absence of each of the n symptoms. Given a finite set $\mathcal{S}$ of syndromes, every clinical picture can be classified in terms of the nearest syndrome in the sense of the Hamming distance of the corresponding bitstrings. It will be clear that this rudimentary formalism has considerable drawbacks for diagnosis. For example, the simple fact that some symptoms are more important than others is not taken into account. Even worse, in many cases not all symptoms will be known, or some symptoms may not be applicable at all. The purpose of this paper is to remedy these shortcomings by generalizing the Hamming distance to finite sets. In abstract terms, both a diagnosis and a clinical picture of a patient can be characterized by a finite set of symptoms, as will be explained in the next section.

Before giving a metric for finite sets, let us discuss the adequacy of the three conditions (i)–(iii) above for diagnostical purposes. First, observe that the partitioning of $\{0,1\}^n$ into subsets $N_1, \ldots, N_k$ can be done on the basis of any arbitrary real-valued mapping d. Since (i) expresses that the distance of a syndrome to itself is zero, and (ii) that the distance of one syndrome to some other is equal to the distance of the other syndrome to the first, mappings that do not satisfy (i) and (ii) can hardly be taken into consideration. Condition (iii), the triangle inequality, requires some more justification. We shall argue that (iii) is not only important for geometry, but also for classification. In particular, the triangle inequality provides ground for approximation and locality. Let us illustrate this with an example. Consider a set of syndromes $\mathcal{S}$ and a syndrome $S \in \mathcal{S}$ having distance at least l to any other syndrome. Assume the clinical picture of a patient has distance $\frac{1}{10}l$ to S. By the triangle inequality, the distance to any other diagnosis S' is at least $\frac{9}{10}l$, so 9 times the distance to S. This might provide sufficient evidence in favour of S and against any other diagnosis. Without the triangle inequality one could not exclude so easily the existence of other syndromes than S at a short distance. Of course one has always the possibility to compute all distances between a clinical picture and any syndrome, but this is very inefficient and actually against the idea of approximation.

2 Generalizing Hamming distance

First we consider the situation in which not all symptoms are known, or not all symptoms do apply. Assume we have a set S of symptoms. As we do not wish to fix the number of symptoms in advance, we allow S to be countably infinite. We may just as well number the symptoms, which will be done by simply taking $S = \{0, 1, 2, \ldots\}$, the set of natural numbers. Again we assume that all symptoms are boolean-valued. It seems reasonable to represent syndromes and clinical pictures as finite, single-valued, sets of pairs consisting of a symptom and a boolean. Let $\mathcal{S}$ be the set of all finite, single-valued, sets of such pairs. Whenever convenient, we shall use a string denotation for elements of $\mathcal{S}$, for example, the string `uuftutt` denotes $\{(2, \mathtt{f}), (3, \mathtt{t}), (5, \mathtt{t}), (6, \mathtt{t})\}$. Here `u` is to be interpreted as 'undefined', as opposed to the defined values `t` for 'true' and `f` for 'false'. Strings can be made of equal length by postfixing them with `u`'s.

The first try to generalize the Hamming distance to this new situation is by restricting two given strings s_1, s_2 to the positions in which they are both defined, and then computing their Hamming distance as bitstrings. This try fails since, for example, the unequal strings `ft` and `ut` would have distance 0, thus violating condition (i) of the definition of a metric.

The second try is to take the new strings as three-valued strings and count the number of positions in which two strings differ, exactly as in the two-valued case. This try succeeds mathematically in the sense that indeed a metric is obtained, but this metric is unsatisfactory from the point of view of classification. For example, `ut` and `ft` are at the same distance of `tt`, and this is undesirable in a case in which the first symptom in the clinical picture `ut` does not apply to the patient in question. Another drawback is that the distance between `uuuuf` and `uuuut` is exactly the same as the distance between `ttttf` and `ttttt`, whereas in the latter case 4 out of 5 defined values (symptoms) match. It is not so obvious how to norm the metric in such a way that a larger number of matching *defined* values reduces the distance.

For the third try we need the notion of *symmetric difference* $X \triangle Y$ of two sets X and Y. By definition, $X \triangle Y = (X - Y) \cup (Y - X)$, equivalently characterized by $X \triangle Y = (X \cup Y) - (X \cap Y)$. For any finite set Z, let $|Z|$ denote the number of its elements. Define $d(X, Y) = |X \triangle Y|$ for all $X, Y \in \mathcal{S}$. Now d is a metric according to [4, Chapter 10]. Moreover, $d(\mathtt{ut}, \mathtt{tt}) = 1$ and $d(\mathtt{ft}, \mathtt{tt}) = 2$, so d is not as bad as the second try. However, we still have that $d(\mathtt{uuuut}, \mathtt{uuuuf}) = d(\mathtt{ttttt}, \mathtt{ttttf}) = 2$. The idea is now to norm this metric.

The fourth try is $d(X, Y) = |X \triangle Y| \div (|X| + |Y|)$, a normed version of the previous try. Unfortunately this is not a metric, as the triangle inequality fails: $d(\mathtt{ut}, \mathtt{tu}) = 1$ and $d(\mathtt{ut}, \mathtt{tt}) = d(\mathtt{tt}, \mathtt{tu}) = \frac{1}{3}$.

As fifth and final try we propose the metric

$$d(X, Y) = |X \triangle Y| \div |X \cup Y|$$

Now $d(\mathtt{ut},\mathtt{tt}) = \frac{1}{2}$, $d(\mathtt{ft},\mathtt{tt}) = \frac{2}{3}$, $d(\mathtt{uuuut},\mathtt{uuuuf}) = 1$ and $d(\mathtt{ttttt},\mathtt{ttttf}) = \frac{1}{3}$, all in accordance with the desiderata above. In the next section we prove that d is indeed a metric and show how to accomodate weighting of symptoms.

3 Finite sets as a complete metric space

For any two finite sets that are not both empty, define as above

$$d(X,Y) = |X \mathbin{\triangle} Y| \div |X \cup Y|$$

and complete the definition of d by putting $d(\emptyset,\emptyset) = 0$. In this section we prove that d is a metric. Moreover, any non-empty set of finite sets equipped with the metric d forms a *complete* metric space. We include the latter result for the sake of completeness, although we do not claim it to be particularly relevant to classification. For the definition of completeness of a metric space we refer to [5].

By definition, $d(X,Y)$ is a rational number between 0 and 1. We have $d(X,Y) = 1$ if and only if X and Y are disjoint, in particular so if X or Y is empty but not both. Obviously, d is symmetric (since $\cup$ and $\triangle$ are so), and we have $d(X,Y) = 0$ if and only if $X = Y$. In order to prove that d is a metric, it remains to show that d satisfies the triangle inequality $d(X,Z) \leq d(X,Y) + d(Y,Z)$ for all finite sets X, Y, Z. Below we shall use $X \uplus Y$ for the union of *disjoint* sets X and Y. Note that this so-called disjoint union is symmetric and associative. We write $X \uplus Y \uplus Z$ for the union of the sets X, Y, Z when these sets are pairwise disjoint. For the proof of the triangle inequality we need the following lemma.

LEMMA 1. For all sets X, Y, Z we have $(X-Y)\cup(Y-X)\cup(Z-Y)\cup(Y-Z) = ((X\cup Z)-(X\cap Z)) \uplus (Y-(X\cup Z)) \uplus ((X\cap Z)-Y)$

PROOF. Both sides are equal to the set $S = (X \cup Y \cup Z) - (X \cap Y \cap Z)$. The left hand side is equal to S since this set consist of those elements that occur in at least one and at most two of the sets X, Y, Z. The right hand side is equal to S on the basis of the following calculation: $(X \cup Y \cup Z) - (X \cap Y \cap Z) = ((X \cup Z) - (X \cap Y \cap Z)) \uplus ((Y - (X \cup Z)) - (X \cap Y \cap Z)) = ((X \cup Z) - (X \cap Y \cap Z)) \uplus (Y - (X \cup Z)) = ((X \cup Z) - (X \cap Z)) \uplus ((X \cap Z) - Y) \uplus (Y - (X \cup Z))$. $\square$

If one or more of the sets X, Y, Z is empty, then the triangle inequality trivially holds. Now assume that X, Y, Z are finite non-empty sets. Using the elementary fact $|A \uplus B| = |A| + |B|$ (additivity), and its immediate consequences $|A \cup B| \leq |A| + |B|$ and $A \subseteq B \Rightarrow |A| \leq |B|$ as well as the lemma above, we can calculate

$$d(X,Y) + d(Y,Z) \quad = \quad \frac{|(X-Y)\cup(Y-X)|}{|X\cup Y|} + \frac{|(Y-Z)\cup(Z-Y)|}{|Y\cup Z|}$$

$$
\begin{aligned}
&\geq \frac{|(X-Y)\cup(Y-X)\cup(Y-Z)\cup(Z-Y)|}{|X\cup Y\cup Z|}\\
(1)\qquad &= \frac{|((X\cup Z)-(X\cap Z))\uplus(Y-(X\cup Z))\uplus((X\cap Z)-Y)|}{|X\cup Y\cup Z|}\\
&\geq \frac{|((X\cup Z)-(X\cap Z))\uplus(Y-(X\cup Z))|}{|X\cup Y\cup Z|}\\
&= \frac{|((X\cup Z)-(X\cap Z))\uplus(Y-(X\cup Z))|}{|(X\cup Z)\uplus(Y-(X\cup Z))|}\\
(2)\qquad &\geq \frac{|(X\cup Z)-(X\cap Z)|}{|X\cup Z|}\\
&= d(X,Z)
\end{aligned}
$$

The numbered formulas above should be explained a bit more: (1) uses the lemma; (2) uses that $\frac{x+y}{z+y}$ is monotone in $y \geq 0$ for $z \geq x \geq 0$. This concludes the proof of the triangle inequality.

Observe that the only property of $|\,.\,|$ we have used is additivity: $|A \uplus B| = |A| + |B|$. This allows us to accomodate weighting in a very satisfying way. Assume every element x has weight $w(x) > 0$. Redefine $|\,.\,|$ by $|\{x_1, \ldots, x_n\}| = w(x_1) + \cdots + w(x_n)$. Then additivity is preserved and everything above goes through.

We finish this section with the completeness result. Consider a fundamental sequence $X_0, X_1, \ldots$, that is, for every $k > 0$ there exists an N such that $d(X_n, X_m) < \frac{1}{k}$ for all $n, m \geq N$. We have to prove that the sequence $X_0, X_1, \ldots$ has a limit. We will even prove that any fundamental sequence becomes eventually constant. The proof will use the following two lemmas.

LEMMA 2. For all finite sets X and Y, if $d(X,Y) < \frac{1}{2}$, then $|X| \leq 2 \cdot |Y|$.

PROOF. If X and Y are both empty, then the lemma trivially holds. Otherwise, observe $X \cup Y = (X-Y) \uplus (X \cap Y) \uplus (Y-X)$ and put $m = |X-Y|$, $n = |Y-X|$, $k = |X \cap Y|$. Then $d(X,Y) = \frac{m+n}{m+k+n} < \frac{1}{2}$ is equivalent to $m+n < k$, which easily implies $m + k < 2 \cdot (n+k)$, so $|X| \leq 2 \cdot |Y|$. □

LEMMA 3. For all finite sets X and Y of bounded size, $d(X,Y)$ is minimal if either $X \subseteq Y$ or $Y \subseteq X$. Moreover, if $d(X,Y) \neq 0$, then $d(X,Y) \geq \frac{1}{n}$, where n is the maximum of $|X|$ and $|Y|$.

PROOF. One element more in the intersection of X and Y means two elements less in the symmetric difference and one element less in the union. Now the lemma follows immediately since $\frac{x}{x+y}$ is monotone in $x \geq 0$ and antitone in $y \geq 0$. □

Let $X_0, X_1, \ldots$ be a fundamental sequence. By definition there exists N such

that $d(X_n, X_N) < \frac{1}{2}$ for all $n > N$. By Lemma 2 $|X_n| \leq 2 \cdot |X_N|$ for all $n > N$. Hence the sets of a fundamental sequence have bounded size. Applying Lemma 3 it follows that every fundamental sequence is eventually constant.

4 Other applications

The field of genetic algorithms [1] traditionally uses bitstrings to represent genotypes, therefore the Hamming distance is a natural metric to measure distance between these genotypes. Roughly speaking, a genetic algorithm generates repeatedly new individuals, whose bitstrings can be compared with the other individuals of the population. For example, a new individual could replace an old one that is at minimal Hamming distance, thus preserving the genetic diversity of the population.

Genetic *programming* [3] differs from genetic *algorithms* [1] in one important aspect: the 'individuals' are parse trees of computer programs instead of bitstrings. Because of this different structure (involving variable length), the Hamming distance does not apply directly to genetic programming. In [2] first steps are taken to use the generalized Hamming distance introduced in the previous sections to measure distance between parse trees and populations of parse trees.

With the generalized Hamming distance from Section 3 applied to finite sets of subtrees, it is possible to monitor the exploration rate (defined as the amount of change in genetic material between subsequent generations). Moreover, the likelyhood that two parse trees have a common ancestry can be estimated and several strategies can be developed to maintain the structural diversity of the population.

Conclusion

We proposed a distance measure to compare objects that have heterogeneous sets of characteristics, such as, for example, syndromes and clinical pictures in medical diagnosis. We formally proved that the proposed distance measure yields a complete metric space, and argued why in particular the triangle inequality is important. We also mentioned an application in the field of genetic programming.

References

[1] J. Holland. *Adaptation in natural and artificial systems.* MIT Press, 1976.

[2] M. Keijzer. Efficiently representing populations in genetic programming. To appear in *Advances in Genetic Programming*, eds. P.J. Angeline and K.E. Kinnear, MIT Press, 1996.

[3] J. Koza. *Genetic Programming: on the programming of computers by means of natural selection.* MIT Press, 1991.

[4] J.C. Oxtoby. *Measure and category.* Graduate texts in mathematics, 2, 1971.

[5] W.A. Sutherland. *Introduction to metric and topological spaces.* Oxford University Press, 1976.

[6] N.N. Webster. *The new lexicon of the english language.* Lexicon Publications Inc., 1990.