A Multi-Way Divergence Metric for Vector Spaces

Robert Moss and Richard Connor

Department of Computer and Information Sciences,
University of Strathclyde, Glasgow, G1 1XH, United Kingdom
{robert.moss,richard.connor}@strath.ac.uk

Abstract. The majority of work in similarity search focuses on the efficiency of threshold and nearest-neighbour queries. Similarity join has been less well studied, although efficient indexing algorithms have been shown. The multi-way similarity join, extending similarity join to multiple spaces, has received relatively little treatment.

Here we present a novel metric designed to assess some concept of a mutual similarity over multiple vectors, thus extending pairwise distance to a more general notion taken over a set of values. In outline, when considering a set of values X, our function gives a single numeric outcome $D(X)$ rather than calculating some compound function over all of $d(x, y)$ where x, y are elements of X.

$D(X)$ is strongly correlated with various compound functions, but costs only a little more than a single distance to evaluate. It is derived from an information-theoretic distance metric; it correlates strongly with this metric, and also with other metrics, in high-dimensional spaces. Although we are at an early stage in its investigation, we believe it could potentially be used to help construct more efficient indexes, or to construct indexes more efficiently.

The contribution of this short paper is simply to identify the function, to show that it has useful semantic properties, and to show also that it is surprisingly cheap to evaluate. We expect uses of the function in the domain of similarity search to follow.

Keywords: distance metric, multi-way divergence

1 Introduction

Much of the research on similarity search focuses on the similarity (or distance) between two vectors. For many situations, however, ascertaining the mutual similarity of a set of vectors would be useful. Applications could be, for example, similarity joins, clustering, cluster analysis, and potentially many more that are dependent upon these techniques.

The calculation of density, or multi-way divergence as the analogue to distance, has been rarely used and is typically calculated through some compound function over the set of pair-wise distances within the set of vectors: for example, the mean intra-set distance or the mean distance from each vector to a centroid.

In this paper, we derive a function to calculate the divergence of a set of vectors that is based on an existing distance function. This new metric, which is grounded in information theory, avoids the problem of repeated calls to the distance metric through a direct, calculable notion of multi-way divergence without relying on approximation. It reuses the notion of complexity offered in the definition of the original distance metric and reverts to this definition when the size of the set is two. It is bounded, giving a maximum value when there is no commonality, allowing absolute comparisons to be made. And finally, it is only a little more expensive than a single distance to evaluate.

We show that multi-way divergence offers a cheaper alternative to compound functions whilst giving similar, or better, semantic properties. Although we are at an early stage in our investigation, we believe it could be applied usefully to construct new metric indices or to execute more complex queries, such as similarity joins, efficiently.

2 Related work

The notion of a multi-way distance metric is not new, motivation coming from geometry and topology. Recently a few papers have analysed the generalisation to multi-way for any existing metric [3, 5, 6]. They consider whether various axioms are observed in these generalisations. For example, a metric space with a simple dyadic metric has the following axioms:

$$\begin{aligned} d(x,y) &= 0 \Leftrightarrow x = y \\ d(x,y) &= d(y,x) \\ d(x,y) &\leq d(x,z) + d(y,z) \end{aligned}$$

The first of these, is simply generalised to $D(x_1, \ldots, x_n) = 0$ if and only if all x_i are equal; the second to $D(x_{\pi(1)}, \ldots, x_{\pi(n)}) = D(x_1, \ldots, x_n)$ for every permutation π of $\{1, 2, \ldots, n\}$; while the third, triangle inequality, has been generalised in numerous ways [6], such as polyhedron inequality:

$$(n-1) \cdot D(x_1, \ldots, x_n) \leq \sum_{i=1}^{n} D(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_{n+1})$$

These generalised axioms are interesting in their own right, and the first two are desirable properties of a multi-way divergence function, but it is not clear if polyhedron inequality aids similarity search at this stage.

Deza and Rosenberg introduced the multi-way extension of the star distance in [3], while perimeter distance [6] gives a geometrical "average distance". Both of these, however, are compound functions and do not fundamentally look at generalising any specific metric.

3 Structural Entropic Divergence

We consider structural entropic distance (SED) proposed by [2] as a metric over trees, and shown as a vector distance in [1]. This metric operates over probability vectors where the sum of a vector's components equals 1. It is a ratio of the complexity of the mean vector to the geometric mean of complexities of individual vectors, where complexity is defined in terms of Shannon entropy [4] and is the amount of information required to describe the vector.

$$D(x_1, x_2) = \frac{C(\frac{x_1+x_2}{2})}{\sqrt{C(x_1)C(x_2)}} - 1$$

where $C(x) = b^{-\sum_i x_i \log_b x_i}$. They observe that the complexity of the mean vector will be the same as each individual complexity when both vectors are equal; that when the vectors have no intersecting components, the complexity of the mean vector is equal to the sum of individual complexities; and, for any other point, the complexity of the mean vector lies between these bounds.

3.1 Generalisation to a multi-way function

We observe a generalisation of this function to multiple arguments (MSED), while preserving the essence of these properties. Considering the numerator first, we can easily extend to multiple arguments as follows:

$$C\left(\frac{x_1+x_2}{2}\right) \Rightarrow C\left(\sum_{i=1}^{n} \frac{x_i}{n}\right)$$

Now consider the denominator, a geometric mean, this too we extend to the following form:

$$\sqrt{C(x_1)C(x_2)} \Rightarrow \sqrt[n]{\prod_{i=1}^{n} C(x_i)}$$

The ratio of these two terms turns out to give a function in the range $[1, n]$, which can be scaled back into $[0, 1]$:

$$D(x_1, \ldots, x_n) = \frac{1}{n-1} \cdot \left(\frac{C(\frac{1}{n}\sum_{i=1}^{n} x_i)}{\sqrt[n]{\prod_{i=1}^{n} C(x_i)}} - 1 \right)$$

This generalised function remains a ratio of the complexity of a centroid to the geometric mean of its neighbours' complexities, and has the following properties:

- If all vectors are identical it gives 0
- If all vectors are different it gives 1
- All other inputs give a value between these bounds

Consider now the metric space axioms described in Section 2. SED has already been shown to be a proper metric in [2]. For MSED, we have stated that the lower bound is achieved when all elements are the same, so the first axiom holds. Since all operations involved in both the numerator and denominator are commutative, total symmetry holds too. We do not yet know whether polyhedron inequality holds.

4 Evaluation

We compared MSED to three other measures of multi-way divergence, based both on the SED metric and on Euclidean distance. Tuples of 2, 4, 8, 16 and 32 were chosen over randomly generated 5 and 15 dimensional probability vectors.

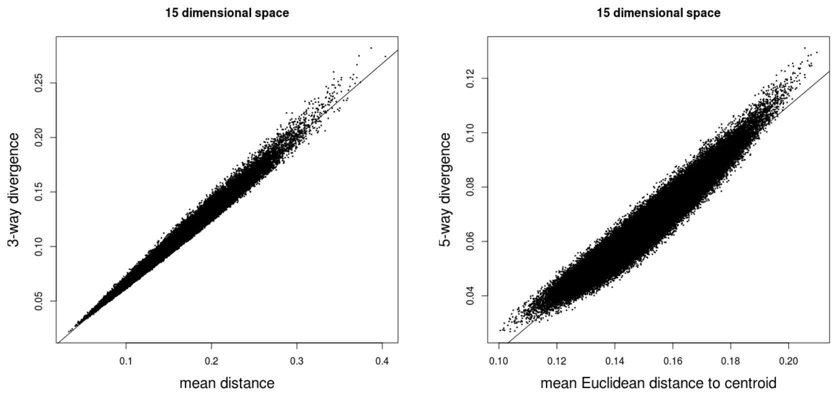
	PCC				
5 dimension SED	2-tuple	4-tuple	8-tuple	16-tuple	32-tuple
mean intra-cluster distance	1	0.9892284	0.9918476	0.9935051	0.9947683
mean distance to a centroid	0.9976659	0.9929966	0.9942112	0.9946998	0.9951615
max intra-cluster distance	1	0.9015272	0.8148037	0.7273526	0.6234713
5 dimension Euclidean					
mean intra-cluster distance	0.9245572	0.9587758	0.966835	0.9741714	0.9770738
mean distance to a centroid	0.9245572	0.9558159	0.9616505	0.964643	0.9684536
max intra-cluster distance	0.9245572	0.8799614	0.7991186	0.7164785	0.6081154
15 dimension SED					
mean intra-cluster distance	1	0.9910392	0.9927492	0.9932659	0.9945081
mean distance to a centroid	0.9979758	0.994636	0.995117	0.9945673	0.9949007
max intra-cluster distance	1	0.8558546	0.764015	0.6177353	0.5374743
15 dimension Euclidean					
mean intra-cluster distance	0.943713	0.9630723	0.9692825	0.9727428	0.97439
mean distance to a centroid	0.943713	0.9618546	0.9676469	0.9700614	0.9718063
max intra-cluster distance	0.943713	0.8333269	0.7624881	0.6385571	0.5282632

Table 1: Pearson's correlation coefficients for MSED

The best correlation comes with the mean distance to a centroid. This method is calculated by making a centroid using the mean vector then averaging all distances in the cluster to it. Since MSED is a ratio of the complexity of a centroid to the geometric mean of individual complexities, there is much more in common with this definition. Even when the comparison distance metric is Euclidean distance, a very strong correlation exists (figure 2b).

The mean intra-cluster distance also correlates well with MSED. Figure 2a shows the correlation with the mean intra-cluster distance, which appears to be strongest at the, more commonly used, lower end.

MSED correlates – less strongly than the other methods – with both SED and Euclidean maximum intra-cluster distance, and the correlation drops as the cluster size increases. Rather than assessing the mutual similarity, the maximum distance simply describes the two farthest points in the cluster. These two points must lie on the cluster perimeter and describe the spread of points across the space. Using only two points, however, fails to account for the spread in other dimensions, further verified by the drop in correlation in the higher dimensional space.



(a) triples with mean structural entropic distance

(b) 5-tuples with mean euclidean distance to a centroid

Fig. 1: Correlation with divergence in 15-dimensional space

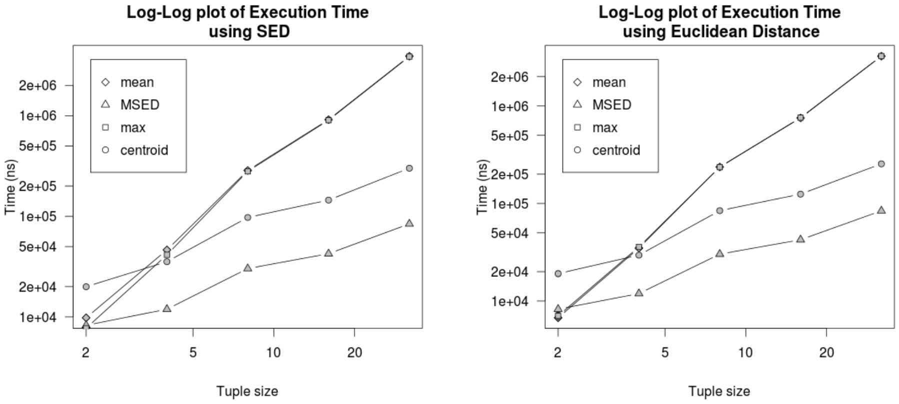
4.1 Performance

When many calculations are performed over a given metric space, the complexity of individual vectors need only be calculated once and stored for later use, thus amortising the cost of the calculation when multiple calls to divergence are required. The only calculation required is the complexity of the centroid, followed by some simple arithmetic, making the calculation really quite cheap.

The performance of the compound functions depend on the number of internal distance calls required: mean distance to centroid is linear, while mean and maximum intra-cluster distances are quadratic. We measured this to compare with MSED, Figure 2 shows the average time in nanoseconds to evaluate each function. While MSED clearly grows linearly it is approximately 3 times faster than mean distance to centroid using euclidean distance, and 4 times faster when using SED.

5 Conclusion

We have shown a formulation of MSED in a single calculation that is based upon information theory; that it is semantically comparable to other more expensive techniques that approximate mutual similarity through averaging, and that the cost to evaluate it is low in comparison with other approximations.



(a) Performance vs SED

(b) Performance vs Euclidean Distance

Fig. 2: Performance

At this point, we find the divergence metric interesting in its own right, and have no very clear idea how it may be usefully deployed. However we have shown that it gives a useful semantic measure of some concept such as *density* within a set of objects, and yet is surprisingly cheap to evaluate compared with other approximations to this concept. We believe that this function will turn out to be useful in the domain of similarity search.

References

1. R. Connor and R. Moss. A multivariate correlation distance for vector spaces. In *Proceedings of the Fifth International Conference on SImilarity Search and APplications*, pages 209–225, 2012.
2. R. Connor, F. Simeoni, M. Iakovos, and R. Moss. A bounded distance metric for comparing tree structure. *Inf. Syst.*, 36(4):748–764, 2011.
3. M.-M. Deza and I.G. Rosenberg. n-semimetrics. *European Journal of Combinatorics*, 21(6):797 – 806, 2000.
4. C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.
5. M. J. Warrens. k-adic similarity coefficients for binary (presence/absence) data. *Journal of Classification*, 26(2):227–245, 2009.
6. M. J. Warrens. n-way metrics. *Journal of Classification*, 27(2):173–190, 2010.