Evaluation of Jensen-Shannon Distance over Sparse Data

Richard Connor[1], Franco Alberto Cardillo[2], Robert Moss[1], and Fausto Rabitti[2]

[1] Department of Computer and Information Sciences,
University of Strathclyde, Glasgow, G1 1XH, United Kingdom
[2] ISTI (Information Science and Technology Institute)
National Research Council of Italy, Via Moruzzi 1, 56124 Pisa (Italy)
{richard.connor,robert.moss}@strath.ac.uk
{franco.alberto.cardillo,fausto.rabitti}@isti.cnr.it

Abstract. Jensen-Shannon divergence is a symmetrised, smoothed version of Küllback-Leibler. It has been shown to be the square of a proper distance metric, and has other properties which make it an excellent choice for many high-dimensional spaces in $\mathbb{R}^*$.

The metric as defined is however expensive to evaluate. In sparse spaces over many dimensions the Intrinsic Dimensionality of the metric space is typically very high, making similarity-based indexing ineffectual. Exhaustive searching over large data collections may be infeasible.

Using a property that allows the distance to be evaluated from only those dimensions which are non-zero in both arguments, and through the identification of a threshold function, we show that the cost of the function can be dramatically reduced.

1 Introduction

Jensen-Shannon divergence is the name given in [8] to a divergence function probably first identified in [10]. It is a simple derivation from Küllback-Leibler [7] yet is positive, symmetric, bounded, and well-defined in the presence of zero values.

Two authors [4, 9] have independently established that one form of Jensen-Shannon divergence is the square of a proper metric. Since then the metric has attracted some more interest in both statistics and information theory, and deeper analysis e.g. [5] shows that it has some properties that, in short, should lend it to being an excellent semantic distance function in many contexts.

The fact that a form exists which is a proper metric immediately leads to the possibility of its use within metric indexing techniques. However many probabilistic spaces are high-dimensional and sparse, and typical Intrinsic Dimensionality [1] is very high: metric indexing techniques are unlikely to be effective.

In this paper we show a way of significantly reducing the cost of similarity search using Jensen-Shannon, showing how an equivalent metric can be derived which requires access only to the intersecting dimensions of the objects being compared. This allows a much more efficient evaluation, and in particular an evaluation which can be performed over inverted indices, thus also subject to parallel evaluation.

2 Definitions and Algebraic Derivations

Jensen-Shannon divergence is defined in terms of Küllback-Leibler divergence:

$$JS(v,w) = \tfrac{1}{2}KL(v,m) + \tfrac{1}{2}KL(w,m)$$

where m is the vector mean of v and w. If logs are taken to base two, then the outcome is bounded in $[\,0{,}1]$.

Some simple algebra gives some other forms of interest for the same function:

$$JS(v,w) = H(m) - \tfrac{1}{2}H(v) - \tfrac{1}{2}H(w) \tag{1}$$

where H is Shannon's entropy function. This can be evaluated as:

$$JS(v,w) = \tfrac{1}{2}\sum_i \big(v_i \log(v_i) + w_i \log(w_i) - (v_i + w_i)\log \tfrac{1}{2}\,(v_i + w_i)\big) \tag{2}$$

From this also can be derived:

$$JS(v,w) = 1 - \tfrac{1}{2}\sum_i \mathcal{F}(v_i, w_i) \tag{3}$$

for a kernel function $\mathcal{F}$ defined by

$$\mathcal{F}(x,y) = h(x) + h(y) - h(x+y)$$

where $h(x) = -x\log_2(x)$.

From this form it may be observed that the evaluation of JS can be achieved with reference only to those dimensions where v_i and w_i are *both* non-zero. A similar form to Equation 3 was given in [3] where the observation was made that this could give an efficient evaluation, but was not quantified.

2.1 Threshold calculation

If the purpose of the distance calculation is as a part of a threshold search, the threshold requirement (using the proper metric form) is:

$$\sqrt{1 - \frac{1}{2}\sum_i \mathcal{F}(v_i, w_i)} < t$$

for threshold t. The function $\mathcal{F}$ can be seen as a similarity accumulator, reaching the value of 2 for perfect similarity, and the term $2t^2$ can be viewed as the maximum shortfall which may occur in order for the threshold t not to be exceeded.

A cost-saving strategy may be used based on this observation. At any point of the iterative calculation, if it can be determined that it is impossible for the value of $\sum_i \mathcal{F}(v_i, w_i)$ to reach the threshold of $2 - 2t^2$, then the calculation may be abandoned.

If stage k of the calculation is considered:

$$\sum_i \mathcal{F}(v_i, w_i) = \sum_{i=1..k} \mathcal{F}(v_i, w_i) + \sum_{i=k+1..n} \mathcal{F}(v_i, w_i)$$

the value of the left hand term is known, and an upper bound for the right-hand term can be found using the Jensen inequality, as $\mathcal{F}$ is a convex function:

$$\sum_{i=j..k} \mathcal{F}(v_i, w_i) \leq \mathcal{F}\left(\sum_{i=j..k} v_i, \sum_{i=j..k} w_i\right)$$

where the value $\sum_{i=k+1..n} v_i$ is simply the complement of $\sum_{i=1..k} v_i$. Therefore, at any stage k of the calculation, the following inequality can be tested:

$$\sum_{i=1..k} (\mathcal{F}(v_i, w_i)) + \mathcal{F}\left(1 - \sum_{i=1..k} v_i, 1 - \sum_{i=1..k} w_i\right) < 2 - 2t^2 \qquad (4)$$

and, if the outcome is true, the final distance calculation will be greater than t.

3 Evaluation

3.1 Definitions

For each test sparse vectors and inverted indices were implemented in a straightforward manner, such that no zero values are stored. Based on the above observations, five different versions of the metric over sparse vector spaces were tested[3]:

Definition 1 An algorithm based on Equation 2, accessing all dimensions of the sparse vectors being compared.

Definition 2 An algorithm based on Equation 3. The algorithm iterates through all nodes of each argument vector, but no calculation is performed if the dimension is not present in both vectors.

Definition 3 The same algorithm as Defn. 2, but at each stage the current accumulator value is checked against a calculated threshold derived from Inequation 4, and the calculation is abandoned when possible.

Definition 4 The accumulation of values is performed over inverted index data structures. The calculation proceeds one dimension at a time, with a separate accumulator being maintained for each object in the set

Definition 5 Again over the inverted index structures, this time maintaing a threshold based on Inequation 4; if the accumulator for any object in the set drops below the required threshold, this is set to -1 and no further calculations are performed over other dimensions of that vector.

The thresholds in Definitions 3 and 5 cause each test to return 10^{-5} of the data.

[3] All implementations are in Java; the source code is available from the authors.

3.2 Framework

For each data set tested, 10^5 separate objects were considered and each one measured against the other members of the set to perform 10^{10} calculations. All code was implemented in Java and executed on a 1.8 GHz Intel Core i7 processor with 4GB of memory; all nonessential processes and network access were disabled. All data structures fitted within the Java heap. Each test was repeated until the standard error of the mean time measured was less than 1%, with a garbage collection being called between each iteration.

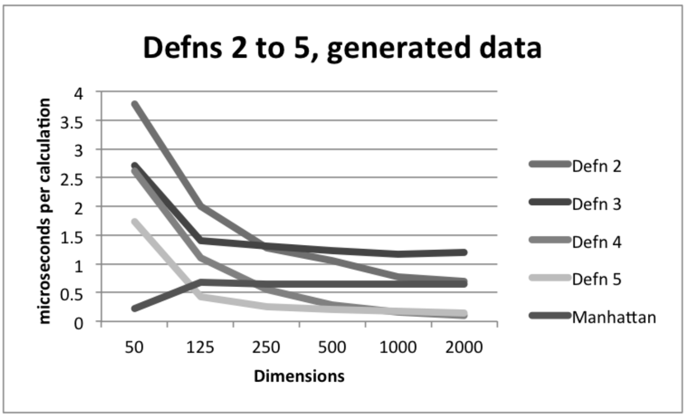

Fig. 1: Cost per calculation using the different implementations

3.3 Generated Spaces

To test the mechanisms over sparse Cartesian spaces a number of generated spaces were used. For each of these, the generator was set to populate a mean of 50 dimensions within all of those available, with the maximum number of dimensions being set between 50 (i.e. a dense space) and 2000.

3.4 Real Spaces

We used the following data sets: *colors*, taken from the colors.ascii file of the SISAP data collection; *english*, taken from the English.dic file of the SISAP collection, from which vectors are generated by the probabilistic technique given in [2]; *occs*, a file of occupations taken from census data using the same generation technique; and *MF-eh*, *MF-ht* taken from the MIR-flickr collection [6]. The key characteristics of these sets is given in Table 1.

Results The results of applying the five techniques to the different data sets are given in Table 2. Definitions 2 to 5 over the generated data are repeated in the graph shown in Figure 1. As well as giving the costs for the five definitions, the cost of Manhattan distance over the sparse representations is also shown.

The results certainly vindicate the techniques described; it is notable that for every truly sparse data set, Jensen-Shannon evaluated by Definition 5 outperforms Manhattan distance, clearly because less data is being moved though the processor. It is equally interesting to note that, even for non-sparse data, the use of inverted indices with a threshold cutoff performs an order of magnitude better than doing per-object comparisons.

The cost-benefit tradeoff for the threshold calculation over generated data is clear to see from Figure 1. While the threshold cutoff is highly effective for some real data sets, it is much less so for others for reasons we do not yet fully understand.

Table 1: Data set characteristics

	colors	english	occs	MF-eh	MF-ht
Total Dimensions	78	483	865	150	43
Mean non-zero dimensions	40.1	16.0	38.3	142.9	43.0
IDIM	5.79	87.6	74.7	5.57	2.37

Table 2: Time (μs) per distance calculation

	Generated Sets						Real data sets				
Implementation	50	125	250	500	1000	2000	colors	english	occs	MF-eh	MF-ht
Defn. 1	3.926	5.325	5.641	6.003	5.952	6.099	3.556	1.838	4.229	11.638	3.391
Defn. 2	3.775	1.995	1.276	1.061	0.772	0.692	2.416	0.410	1.180	10.558	3.335
Defn. 3	2.711	1.404	1.307	1.232	1.162	1.195	1.396	1.139	1.157	2.158	1.934
Manhattan	0.222	0.675	0.652	0.655	0.651	0.649	0.256	0.206	0.429	0.646	0.197
Defn. 4	2.622	1.109	0.550	0.286	0.159	0.103	1.631	0.185	0.603	7.420	2.275
Defn. 5	1.739	0.422	0.253	0.208	0.178	0.140	0.205	0.152	0.199	1.127	0.260

4 Conclusions and Further Work

In this paper we have shown how two algebraic deductions from the Jensen-Shannon distance can be used to give a very significant cost saving in its evaluation. The inverted index implementation is also perfectly suited to parallelisation, and in particular can use parallel threads on a graphics accelerator rather than specialist hardware. In combination, we believe this metric is made much more accessible.

We have not yet fully investigated the threshold cutoff. In particular, the results shown here evaluate the threshold at every stage of the algorithms, which applies a significant cost for little benefit at the early stages. Different collections behave in different ways, but it should be easily possible to determine a better strategy when the calculation is made only when there is a significant chance of aborting the calculation.

Acknowledgements

The authors would like to acknowledge funding from the UK Research Councils for the related project EP/G012407/1. Richard Connor was supported by the National Research Council of Italy (CNR) for a Short-term Mobility Fellowship (STM), which funded a stay at ISTI-CNR in Pisa where much of this work was done.

References

1. Edgar Chávez, Gonzalo Navarro, Ricardo Baeza-Yates, and José Luis Marroquín. Searching in metric spaces. *ACM Comput. Surv.*, 33(3):273–321, September 2001.
2. Richard C. H. Connor, Fabio Simeoni, Michael Iakovos, and Robert Moss. Towards a universal information distance for structured data. In Alfredo Ferro, editor, *SISAP*, pages 69–77. ACM, 2011.
3. Ido Dagan, Lillian Lee, and Fernando C. N. Pereira. Similarity-based models of word cooccurrence probabilities. *Mach. Learn.*, 34(1-3):43–69, February 1999.
4. D.M. Endres and J.E. Schindelin. A new metric for probability distributions. *Information Theory, IEEE Transactions on*, 49(7):1858–1860, 2003.
5. B. Fuglede and F. Topsoe. Jensen-shannon divergence and hilbert space embedding. In *Information Theory, 2004. ISIT 2004. Proceedings. International Symposium on*, pages 31–, 2004.
6. Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*, New York, NY, USA, 2008. ACM.
7. S. Küllback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.
8. Jianhua Lin. Divergence measures based on the shannon entropy. *Information Theory, IEEE Transactions on*, 37(1):145–151, 1991.
9. F. Österreicher and I. Vajda. A new class of metric divergences on probability spaces and and its statistical applications. *Ann. Inst. Statist. Math.*, 55:639–653, 2003.
10. C. Radhakrishna Rao. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, 44(1):pp. 1–22, 1982.