

Sequence Distance Embeddings

Graham Cormode, grahamc@dcs.warwick.ac.uk

(joint work with Mike Paterson, S. Muthukrishnan,
Cenk Sahinalp, Uzi Vishkin, Piotr Indyk,
Mayur Datar, Nick Koudas)

- Vector distances, streaming models
- Biological distances on permutations
- Editing distances on strings

Vector Distances

Two vectors $a[1 \dots n]$ and $b[1 \dots n]$. Distance between them:

- Euclidean, $(\sum (a[i] - b[i])^2)^{1/2}$
- Manhattan, $\sum |a[i] - b[i]|$
- Hamming, $\sum (a[i] \neq b[i])$

Generalise this to the L_p distance =

$$\|a - b\|_p = (\sum |a[i] - b[i]|^p)^{1/p}$$

Euclidean = L_2 , Manhattan = L_1 , Hamming $\sim L_0$ (as $p \rightarrow 0$, $L_p \rightarrow$ Hamming). Also if $a[i], b[i] \in \{0,1\}$ then $(L_p)^p = \text{Hamming}$.

Note, L_p may be interesting for (fractional) p , $0 < p < 2$

Stable Distributions

Let X be a random variable distributed with a *stable distribution*. Stable distributions have the property that

$$a_1X_1 + a_2X_2 + a_3X_3 + \dots + a_nX_n \sim \|(a_1, a_2, a_3, \dots, a_n)\|_p X$$

if $X_1 \dots X_n$ are stable with stability parameter p

Gaussian distribution is stable with parameter 2

Stable distributions exist and can be simulated for all parameters $0 < p < 2$.

So, let $X = x_{1,1} \dots x_{m,n}$ be a matrix of values drawn from a stable distribution with parameter p ...

Dimensionality Reduction of Vector Problems

$$(a_1 \dots a_n) \bullet \begin{pmatrix} (x_{1,1} \dots x_{m,1}) \\ \dots \\ (x_{1,n} \dots x_{m,n}) \end{pmatrix} = (s_1, \dots s_m) \text{ [a sketch, s]}$$

$$(b_1 \dots b_n) \bullet \begin{pmatrix} (x_{1,1} \dots x_{m,1}) \\ \dots \\ (x_{1,n} \dots x_{m,n}) \end{pmatrix} = (t_1, \dots t_m) \text{ [a sketch, t]}$$

Then $\text{median}(|s_1 - t_1|, |s_2 - t_2|, \dots, |s_m - t_m|) / \text{median}(X)$
is an estimator for $\|a - b\|_p$

Can guarantee the accuracy of this process: will be within
a factor of $1+\epsilon$ with probability δ if $m = O(1/\epsilon^2 \log 1/\delta)$

Streaming model

Suppose n is huge (Mb, Gb, Tb, Pb), vectors are too big to store

Data arrives as a *stream*, eg

$a[3] = 1027$, $a[19] = 76$, $a[61250] = 39$, $a[14235261] = 123$,
 $a[27] = -4369$, $a[41242] = 2$, $a[2] = 1241$, $a[213] = -16...$

The stream may be ordered, or may be in no particular order.
Missing entries are assumed to be 0.

May also be unaggregated eg

$a[3] \leftarrow +12$... $a[3] \leftarrow +7$... $a[3] \leftarrow -5$

Overall effect here is $a[3] = 14$

Sketching the stream

Can compute these sketches on the stream because of linearity of the sketch function, eg

$$a[j] \leftarrow +12$$

$$\text{so } s_1 = s_1 + 12x_{1,j} \quad s_2 = s_2 + 12x_{2,j} \quad \dots \quad s_n = s_n + 12x_{n,j}$$

So just add the right multiple of column j when j is updated

But... we need to store all $x_{i,j} = O(mn)$ storage

... which is more than just storing the vector a

Can we get round this?

Reducing space needs

- $x_{i,j}$ must be from stable distribution with fixed parameter
- $x_{i,j}$ must be the same every time it is used

We cheat a little by using pseudo-randomness.

We will use function **stable** to generate values from a stable distribution from a uniform distribution (like Box-Mueller transform for Normal dbn).

We use a random number generator that is good enough so that $f(x)$ appears to be drawn from a uniform dbn.

Then $x_{1,j} = \text{stable}(f(x))$ $x_{2,j} = \text{stable}(f(f(x)))$
 $x_{3,j} = \text{stable}(f(f(f(x))))$ etc.

Practical Use

So with $O(m)$ space we can create a sketch to allow rapid comparison of huge streaming vectors.

Note $m \ll n$, in fact m is independent of n .

Implemented and tested in:

[C, Indyk, Koudas, Muthukrishnan02] - On massive tabular data, looking for clusterings using sketch computations to speed up comparisons for L_1 , L_2 and other L_p distances

[C, Datar, Indyk, Muthukrishnan] - On streaming vectors, to count number of distinct elements, find Hamming norm and Hamming distance.

Other distance measures

What about other more challenging distance measures eg

- string distances based on alignments
- computational biology measures

Outline approach:

