Measuring Pipeline Diversity

Data Machines Corporation

May 25, 2020

If a pipeline is sequence of primitives, how do we quantify the diversity of a collection of pipelines, in terms of its primitives? We have to decide what we mean by diversity. Is a collection of only two pipelines very diverse if all their primitives are different? How about a collection of many pipelines that differ in only one primitive—is that more or less diverse?

The use the Levenshtein edit distance between pairs of pipelines gives us a number describing the difference (diversity) between the pairs of pipelines, in terms of primitives.

What about a collection? We have a different edit distance for each of N pairs in the collection. We put these N numbers into a vector v of length N—the quantity N is related to the number of pipelines, n, by $N = \frac{(n)(n-1)}{2}$.

As a first idea, we can take the sum or average of the components of this vector. The sum is simply the L_1 norm (see below) because all components are non-negative. We can generalize this to the L_p norm for any p.

$$L_p(v) = \left(|v_1|^p + |v_2|^p + \ldots + |v_N|^p\right)^{\frac{1}{p}}.$$

Each component v_i represents the edit distance between a single pair of pipelines; since this quantity is always positive, the absolute values are not necessary.

What about generalizing the averages? How about:

$$M_p(v) = \frac{L_p(v)}{N}.$$

The quantities $L_p(v)$ and $M_p(v)$ are is defined for all $0 < p < \infty$ but $L_p(v)$ satisfies the properties of a norm only for $p \geq 1$. Two limits can be considered for L and by extension M (the second is not really a limit nor is it a norm, but is often used).

$$L_\infty(v) = \max\{|v_1|, |v_2|, \ldots, |v_N|\}$$

$$L_0(v) = \text{the number of non-zero elements of } v.$$

1 Synthetic Scenarios

Now we test these measures on synthetic data. Consider the following scenarios. All pipelines below have 10 unique primitives.

1. Performer X submits only 2 pipelines—differing by 10 substitutions.
2. Performer Y submits 6 pipelines all differing from each other by 10 substitutions.
3. Performer W submits 2 sets of 3 identical pipelines differing by 10 substitutions between groups.
4. Performer Z submits 5 identical pipelines plus 1 additional pipeline differing from all the rest by 10 substitutions.
5. Performer U submits a sequence of pipelines differing from each other by a number of substitutions equal to the absolute value of the relative position in the sequence.
6. Performer V submits a sequence of pipelines all differing from each other by a single substitution.

2 Relative Orderings

$$L_1 : Y > W > Z > U > V > X \tag{1}$$
$$L_2 : Y > W > Z > U > X > V \tag{2}$$
$$L_3 : Y > W > Z > X > U > V \tag{3}$$
$$L_\infty : Y = W = Z = X > U > V \tag{4}$$
$$L_0 : Y = U = V > W > Z > X \tag{5}$$
$$M_1 : X = Y > W > Z > U > V \tag{6}$$
$$M_2 : X > Y > W > Z > U > V \tag{7}$$
$$M_3 : X > Y > W > Z > U > V \tag{8}$$
$$M_\infty : X > Y = W = Z > U > V \tag{9}$$
$$M_0 : X = Y = U = V > W > Z \tag{10}$$

3 Discussion

Notice that all 10 measures agree that:

$$Y \geq W \geq Z \geq U \geq V$$

If you look back at the scenario, this ordering makes sense.

Now notice that the position of X is very variable throughout the measures. What is different about X? Notice that the collections $Y, W, Z, U,$ and V all had six pipelines whereas X had but two. We can surmise that the differences in the measures appear largely across collections of different sizes.

Note that the following orderings don't make sense:

$$X > Y$$

$$X > W$$

$$X > Z$$

Why? Because all of Y, W, and Z contain a pair of distance 10 pipelines, plus additional pairs at nonzero distance whereas X contains just one pair at distance 10. Clearly X should be deemed less diverse than Y, W, and Z. That said, all 5 of the M measures break at least one of these orderings, whereas none of the L measures do.

Of the 5 remaining measures L_∞ and L_0 both lead to a lot of equalities which may not be desirable. The other three measures all place X in a different relative position to $U > V$. The higher the value of p, the more collections of fewer pipelines that are more distant are deemed more diverse than collections with more pipelines that are separated but less distant.

4 Counterexample

Because all same-sized collections chosen for the examples above have a consistent ordering under the norms, it becomes natural to wonder if same-sized collections are always ordered the same way regardless of the norm used.

This property does *not* hold as the following counterexample shows:

Performer S submits k identical pipelines plus one additional pipeline at distance d. There are $k+1$ pipelines and $k(k+1)/2$ pairs, of which k pairs have distance d and the rest have distance 0. S is the same as Z with $k = 5$ and $d = 10$.

Performer T submits n pipelines and N pairs, where $N = n(n-1)/2$. All of these pipelines have distance e from each other. S and T have the same numbers of pipelines/pairs if $n = k+1$. T is the same as V with $e = 1$, $n = 6$, and $N = 15$.

Both S and T have simple formulas for the measures, coming from their simple structures. The formulas are displayed in the following table:

performer	L_1	L_2	L_∞
S	dk	$d\sqrt{k}$	d
T	eN	$e\sqrt{N}$	e

Filling in the values of these measures for the parameters that define Z and V, we find

performer	L_1	L_2	L_∞
Z	50	22.36	10
V	15	3.87	1

This table reproduces the consistent ordering found above.

Now reparameterize so that $d = 10$, $k = 10$, $e = 2$, $n = 11$, and $N = 55$. Then the values for these measures transform to:

performer	L_1	L_2	L_∞
S	100	31.62	10
T	110	14.83	2

Note the measures L_1 and L_2 reversed the ordering of the diversity of collections S and T.

5 Asymptotics

We can define two parameters to apply to *every* collection of pipelines: let d, for *diameter*, denote the largest edit distance across all pairs. And let n denote the *number* of pipelines.

In the above section, we defined two collections of pipelines S, and T. Collection S had parameters d and k, and collection T had parameters e and n (or equivalently N). The definitions coincide where the letters coincide. Additionally, bridging notation, k coincides with $n-1$, and e coincides with d.

Collections S and T have significance for asymptotics.

For given values of d and n, the collection T has the most diversity (or there could be ties, but none have greater diversity). Why? Because d is the largest edit distance and all pairs of T have edit distance d.

On the other hand, for given values of d and n, the collection S has the least diversity (or again, there could be ties, but none have less diversity). Why? At least one pair must have distance d. Collection S has one pair at distance d, as required, then makes all other pipelines identical to just one of these two maximally distant ones. So there are $(n-1)(n-2)/2$ identical pipelines and $n-1$ at maximal distance.

Collection S would seem to be the least diverse, but there is something to prove here. If there are 3 pipelines, a, b, and c, where a and c are at maximal distance d, and the other pairs are at less or equal distance, is it really true that minimal diversity is acheived by placing b at a or by placing b at c, leaving two pairs with distance d, and one pair with distance 0?

Note that (for L_1 minimality) we want to show that for all pipelines b:

$$d_{\text{edit}}(a, b) + d_{\text{edit}}(b, c) \geq d$$

But that is the triangle inequality applied to the edit distance, which can be proved by considering editing a to b and then b to c, which takes you from a to c with no less than d edits.

What about for the other norms, $p \neq 1$? I *think* it is obvious that

$$d^p_{\text{edit}}(a, b) + d^p_{\text{edit}}(b, c) \geq d^p,$$

and the result should follow, but I'll come back to this.

Now let's collect and review the results in the new notation placing bounds on an arbitrary collection C with parameters d and n.

$$d\sqrt[p]{n-1} \leq L_p(\mathrm{C}) \leq d\sqrt[p]{(n)(n-1)/2}$$

Equality hold for the extreme cases, S and T, for lower and upper bounds, respectively.

Put more compactly,

$$d \times O\left(\sqrt[p]{n}\right) \leq L_p(\mathrm{C}) \leq d \times O\left(\sqrt[p]{n^2}\right) \text{ as } n \to \infty$$

Here d and n are positive integers and, moreover, $n \geq 2$ (otherwise there would be no pairs).

The bounds coincide for $n = 2$, equaling d, because there is but one equivalence class of collections (two pipelines at distance d). In these collections, all norms equal d.

For a given value of d, and $n > 2$, the upper bound stays above the lower bound, and for both, the norms are ordered in a consistent way, decreasing with increasing p. Recall, that S and T were on the lower and upper bound, respectively. The counterexample worked for S and T only because the two values of d were different: $10 = d \neq e = 2$ (both $n = 11$, though).

6 Visualizations (coming soon)

It would be informative to add a few visualizations. There are upper and lower bounds and these bounds can change for all values of p, say 1, 2, 3, and ∞. The plots will be shown as a function of n (horizontal axis), the value of d can set the units for the vertical axis – there is no other dependence on d. It isn't clear how many plots we need, but I will show those plus a separate one for the counterexample.

7 Conclusion

Consistent with the story we saw above, higher values of p deem fewer larger edit distances as more diverse than a greater number of smaller edit distances. On the other hand, the reverse is true for lower values of p. Thus, jiving with intuition, there is no one single measure of diversity. Perhaps we can report L_1, L_2, and L_∞ as three relevant measures describing diversity. All of these measures have natural interpretations. Many times, all three proposed measures will agree, as it was hard to find the counterexample above.

If I had to pick one measure to use consistently, without the others, it would be L_2 because it is a natural balance between the two extremes. This is the unique norm whose tight upper bound grows linearly with n.