A Note on the Ground Distance for Weisfeiler-Lehman, Tree Metrics, and the French Railway Metric

Pascal Welke

June 1, 2021

I will be using the notation from your second presentation (draft). In particular, I assume that we have a database $\mathcal{G}$ of labeled graphs $G = (V, E, l_0)$ and the set of Weisfeiler-Lehman (WL) labels of iteration i is denoted by Σ_i.

Distance Between WL Labels with Identical Parents

Now, we want to define a distance function on a *subset* of the pairs from Σ_i. To this end, let's represent each depth-i WL label $\sigma \in \Sigma_i$ for $i \geq 1$ as

$$\sigma = (l_{i-1}(v), \{l_{i-1}(u) | u \in \mathcal{N}(v)\})$$

for some arbitrary representative vertex $v \in V(G)$ for at least one $G \in \mathcal{G}$.

We now define, for all $\sigma_{i-1} \in \Sigma_{i-1}$

$$\mathit{children}(\sigma_{i-1}) = \{\sigma \in \Sigma_i | \sigma = (\sigma_{i-1}, X)\} = \{\{(\sigma_{i-1}, X_1), (\sigma_{i-1}, X_2), \ldots, (\sigma_{i-1}, X_{c_{\sigma_{i-1}}})\}\,.$$

For $(\sigma_{i-1}, X_a), (\sigma_{i-1}, X_b) \in \mathit{children}(\sigma_{i-1})$, we can now finally define

$$d(\sigma_{i-1}, X_a), (\sigma_{i-1}, X_b)) = d_{Hamming}(X_a, X_b) = \frac{1}{m}|X_a \cap X_b|\ ,$$

where $m = \max_{\sigma_{i-1}, X) \in \mathit{children}(\sigma_{i-1}} |X|$ or

$$d'(\sigma_{i-1}, X_a), (\sigma_{i-1}, X_b)) = d_{Jaccard}(X_a, X_b) = 1 - \frac{|X_a \cap X_b|}{|X_a \cup X_b|}\ .$$

For elements $l_a, l_b \in \Sigma_0$, you can use the discrete metric

$$d(l_a, l_b) = \begin{cases} 1 & \text{if } l_a = l_b \\ 0 & \text{if } l_a \neq l_b \end{cases}$$

I hope this makes things more clear.

French Railway Metric

To obtain edge weights for the global WL label tree, we somehow need to transform the $\binom{k}{2}$ pairwise distances between k children of any σ_{i-1} to k distances that we can write to the edges between σ_{i-1} and its children. To this end, we use the french railway metric idea. I suggest you find *a sensible*[1] average or mean point of $children(\sigma_{i-1})$ and use this as Paris. Then, $d(X, \text{Paris})$ (resp. $d'(X, \text{Paris})$) is the weight for the edge between σ_{i-1} and (σ_{i-1}, X).

Wasserstein Distances Between Graphs

While I'm at it, here it goes. Now that you have a tree with weighted edges, you can use the corresponding tree metric to define distances between labels of different WL depths and between labels that are not children of the same parent. You may use this tree metric as ground distance[2] for the Wasserstein distance between the multisets / histograms of WL labels of graphs.

[1] The Euclidean mean is particularly suited for Euclidean metric spaces, as the name suggests. There might be better means for Hamming and Jaccard distances.

[2] So now we have Wasserstein distances based on tree metrics based on french railway metrics based on distances between WL labels with identical parents. :)