Supplementary Materials - A Method for Discovering Novel Classes in Tabular Data

First Author*, Second Author*, Third Author*
* *Affiliation*
email@email.com

I. Comparison of the different pseudo-label definition methods

A. Hyper-parameter range

In order to compare the robustness and efficiency of the threshold and top k based pseudo labelling methods, TabularNCD is evaluated with both methods across all the possible values they can take.

The threshold based method is defined as:

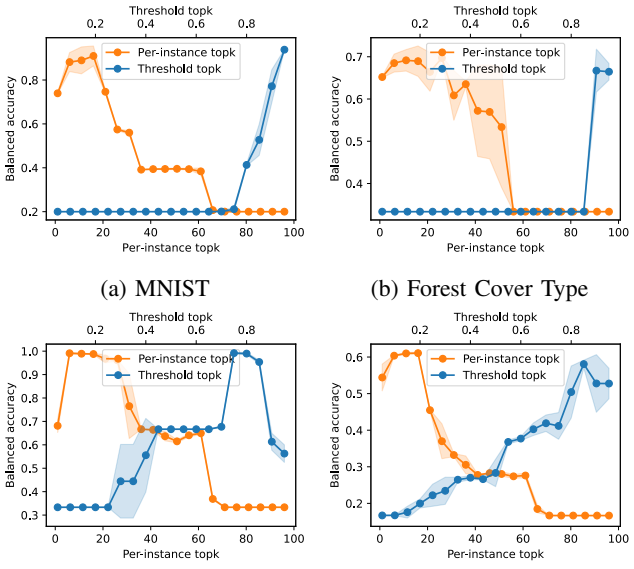
$$\hat{y}_{i,j} = \mathbb{1}[\delta(z_i, z_j) \geq \lambda] \tag{1}$$

where λ is the threshold and $\delta(z_i, z_j)$ is the cosine similarity of z_i and z_r. So this method is evaluated for $\lambda \in]0, 1[$.

And the method based on the number of pairs is defined as:

$$\hat{y}_{i,j} = \mathbb{1}[j \in \underset{\substack{r \in \{1,\dots,|Z|\} \\ r \neq i}}{\text{argtop}_k} \ \delta(z_i, z_r)] \tag{2}$$

where argtop_k is the subset of indices of the k largest elements. So this method is evaluated for $k \in [1, |Z^u| - 1]$, where $|Z^u|$ is the number of unlabelled observations in the mini-batch Z.



(a) MNIST (b) Forest Cover Type

(c) Human Activity Recognition (d) US Census 1990

Fig. 1: Balanced accuracy w.r.t to the pseudo labelling method, for all of their possible values.

The results of this experiment are displayed in Fig 1 and are averaged over 10 executions for 4 different datasets. They reveal that the threshold based method obtains similar or slightly worse performance than the top k based method of Eq. 1, and on a smaller range of values, which makes hyper-parameter tuning more difficult.

B. Clustering collapse

During the experiments, we also noticed that with the threshold based definition, the clustering network would sometimes *collapse*, where it degenerated to a trivial solution and predicted the same class for all instances (see Fig. 2). This issue comes from the fact that the cosine similarity is measured in the latent space, which is updated during learning. The model can therefore minimize the loss by concentrating all of the unlabeled data so close together that the cosine similarity between any two points will be higher than λ, resulting in a trivial solution where any pair of points will be regarded as positive with Eq. 1.

This phenomenon cannot happen with the method of Eq. 2, since the number of positive and negative pairs is not based on the latent representation of the data.

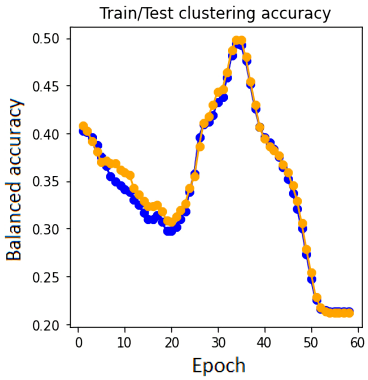


Fig. 2: Evolution of the balanced accuracy during training, where the clustering collapses.