Transferable Deep Metric Learning for Clustering

Mohamed ALAMI [1,2] Jesse READ [1] Rim KADDAH [2]

[1]LIX, Ecole Polytechnique, Ecole Polytechnique, Institut Polytechnique de Paris, 91120 Palaiseau, France

[2]IRT SystemX, Paris-Saclay, France

Due to the curse of dimensionality, clustering in high dimension spaces remains a hard task mainly because distance-based algorithms like k-means are no longer tractable or effective. Moreover, the choice of the metric is crucial as it is highly dependent on the dataset characteristics; Euclidean and other standard distance metrics may not be appropriate. We propose a framework for learning a transferable metric. Using a graph auto-encoder, we show that it is possible to build dataset independent features characterising the geometric properties of a given clustering. These features are used to train a critic that serves as a metric which measures the quality of a clustering. We learn and test the metric on several datasets of variable complexity (synthetic, MNIST, SVHN, omniglot) and achieve close to state of the art results while using only a fraction of these datasets and shallow networks. We show that the learned metric is transferable from a dataset to another even when changing domain or task.

Objective

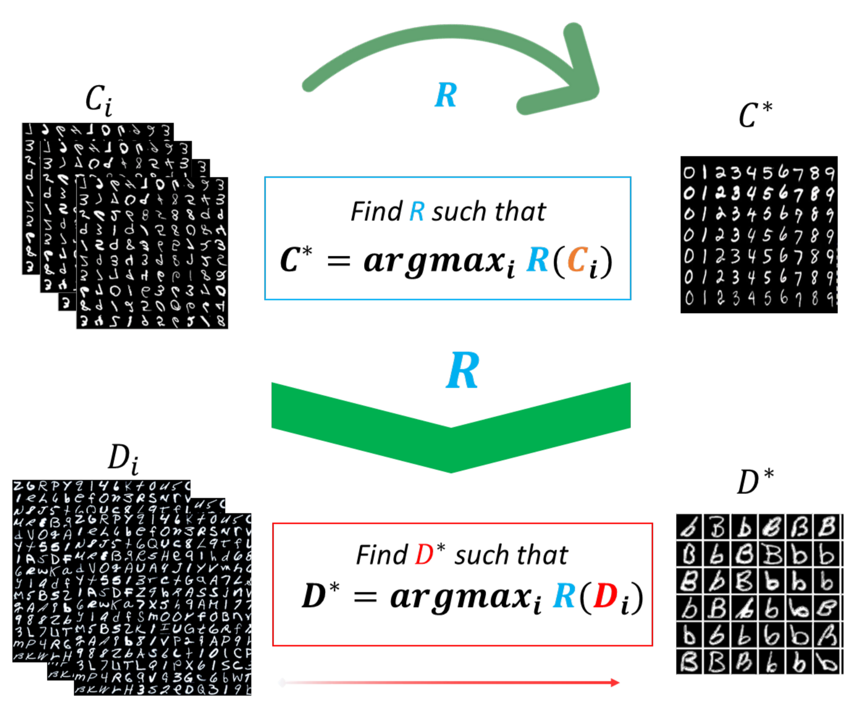

Figure 1:Using available correctly clustered datasets, we find a reward function R that allows to replicate the given clustering and use it to cluster new datasets.

A- Clustering Network

Given a reward function R and a non clustered dataset, a neural network is trained to find the clustering (state) that maximizes R. R grades the quality (value) of the state, therefore states are not independent from each others

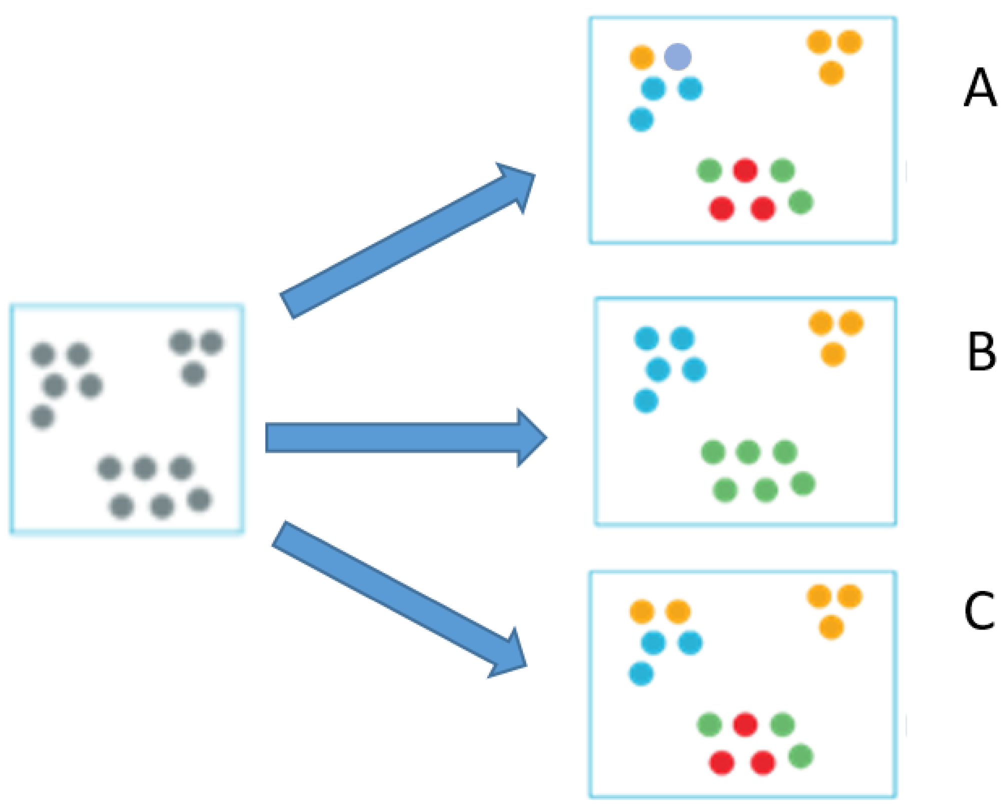

Figure 2:We suppose that R should have the following characteristics:

- $d(A, C) < d(A, B)$ and $B =_i R_i$
- R might not be differentiable.

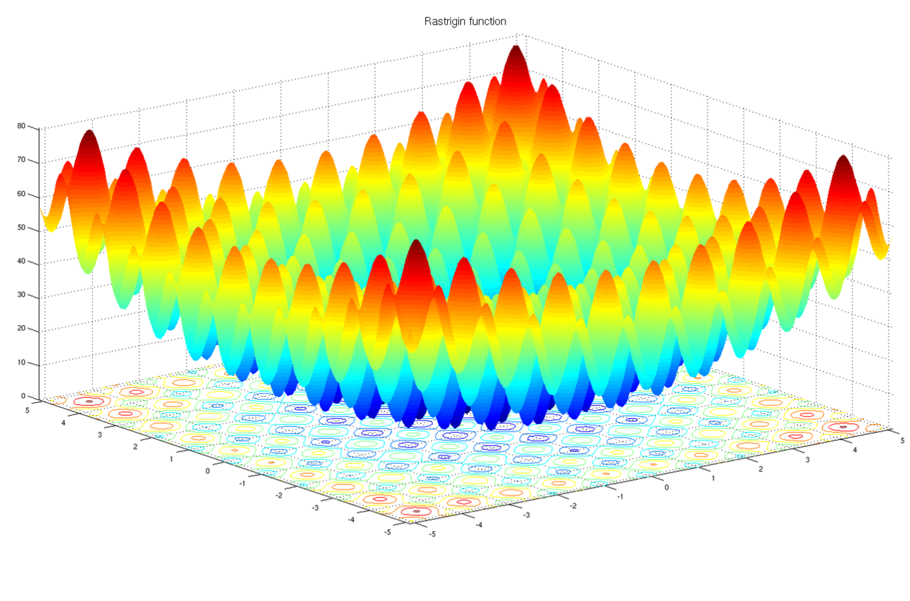

Figure 3:Possible illustration of R in 3D space

Evolutionary algorithms are therefore a good strategy to find the clustering that maximizes R.

B- Graph Based Dataset Embedding

Important Notice

The R metric does not score the similarity between instances in a dataset but the quality of the entire clustering. Moreover the R function has to be transferable between datasets. It is therefore necessary to encode the general geometric properties of a clustering.

The best clustering found by the clustering network is turned into a graph by drawing edges between points of the same cluster. The resulting graph is input into a graph auto-encoder [?] to produce an embedding vector $\hat{z}$ of the clustering (see part B in figure 1).

C- A critic as a metric

The proposed embedding $\hat{z}$ and the target z are input into a Wassertein GAN critic [?] that produces a continuous distance between the true embeddings and the proposed ones (see part C of figure 1).

Complete Framework

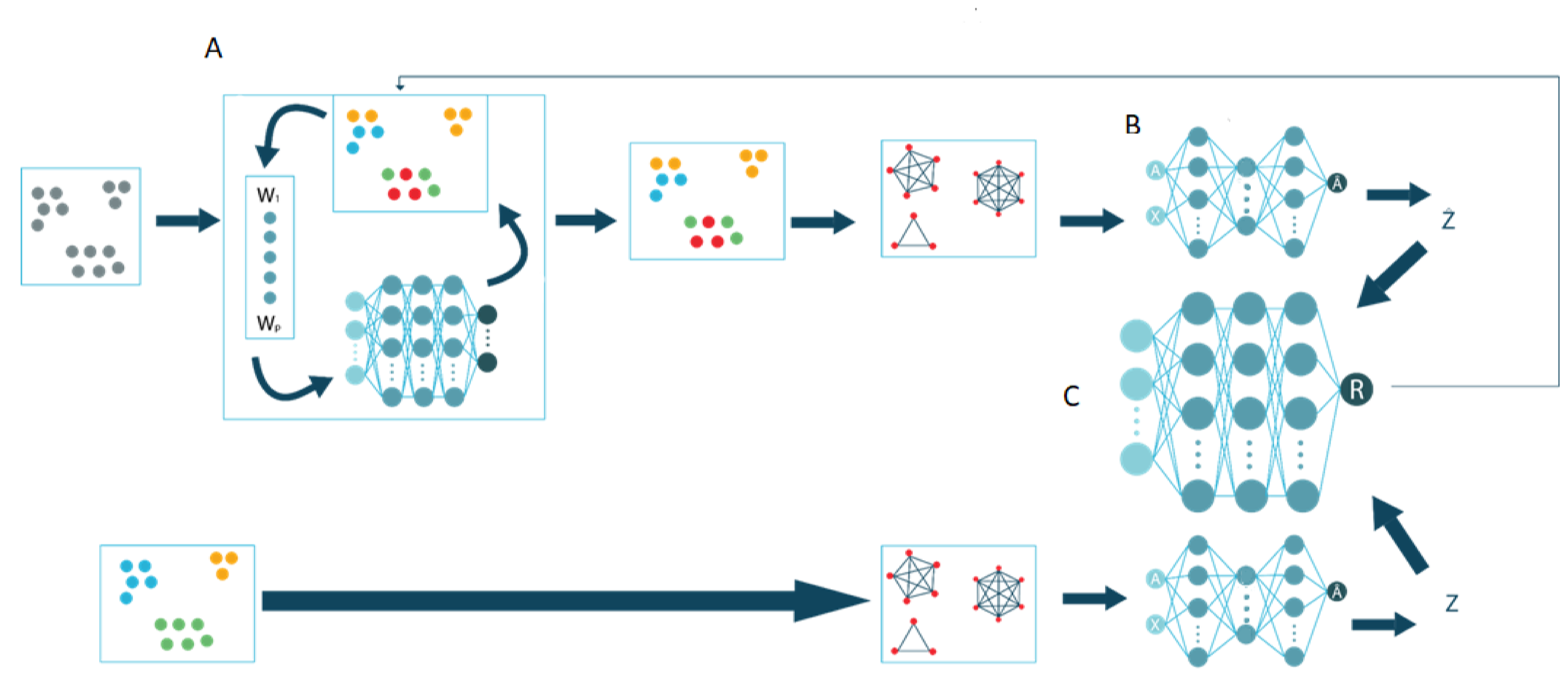

Figure 4:The framework is composed of 3 blocs:

- A: an evolutionary algorithm based agent that performs clustering by maximizing a given reward function
- B: a graph autoencoder that transforms a clustering into a graph then outputs an embedding vector of the graph
- C: a critic neural network that learns a metric function

WGAN is implemented in order to solve the following problem:

r reward function (critic)

$\mathcal{D}$ a set of solutions (i.e. clustering proposals) found using r; $[\mathcal{S}^*]$ the perfect clustering; $[d^*]$ the best solution found in $\mathcal{D}$

The problem becomes:

$$\min_{d^*}\{r(\mathcal{S}^*) - \max_{r} \min_{d \in \mathcal{D} \setminus d^*} r(d^*) - r(d)\} \quad (1)$$
$$\text{s.t} \quad \mathcal{S}^* =_{s \in \mathcal{S}} r(s)$$

Testing Methodology

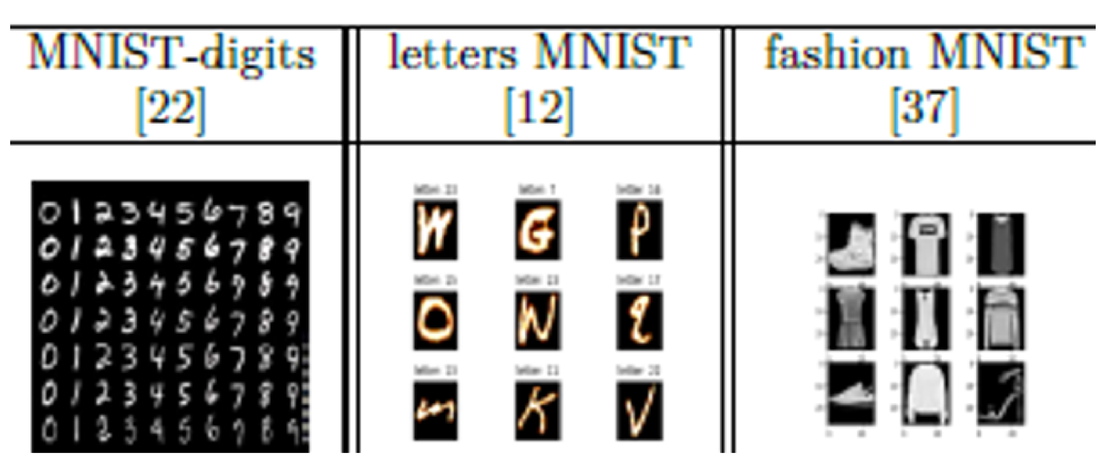

Figure 5:3 MNIST datasets: Numbers, letters ad fashion. In each case, the metric is learned on one dataset then tested on the others.

Results

Training Dataset	Testing Dataset					
	Numbers		Letters		Fashion	
	Best	Top 3	Best	Top 3	Best	Top 3
Numbers (standard)	78.3%	92.5%	86%	97.5%	69.2%	87.2%
Numbers (few shots)	75.8%	82.1%	83.3%	92.0%	65.1%	83.9%
Fashion (standard	70.1%	83.1%	85.0%	98.6%	76.9%	94.7%
Fashion (few shots)	67.9%	77.4%	83.5%	95.3%	70.2%	88.0%

Critic based performance assessment: Best corresponds to the percentage of times the critic gives the best score to the desired solution. Top 3 is when this solution is among the 3 highest scores.

Method	ACC	NMI
CCN [?]	78.18%	0.874
Ours (standard)	83.4%	0.891

When the number of clusters is not known, we outperform the state of the art

References

[1] Thomas N. Kipf and Max Welling, Variational Graph Auto-Encoders, 2016

[2] Martin Arjovsky and Soumith Chintala and Léon Bottou, Wasserstein GAN

[3] Yen-Chang Hsu, Zhaoyang Lv, Zsolt Kira, Learning to Cluster in order to transfer across domains and tasks, ICLR 2018