Dimensionality reduction for clustering with deep neural networks

Agustín Fernández Felguera

Degree in Statistics

01-09-2020

Contents

Introduction

Topic of the project

- Sheer increase in the number of data and its complexity
- Importance of DR and clustering
- Combination of algorithms
- New approaches (Deep learning)

Clustering strategies using DR

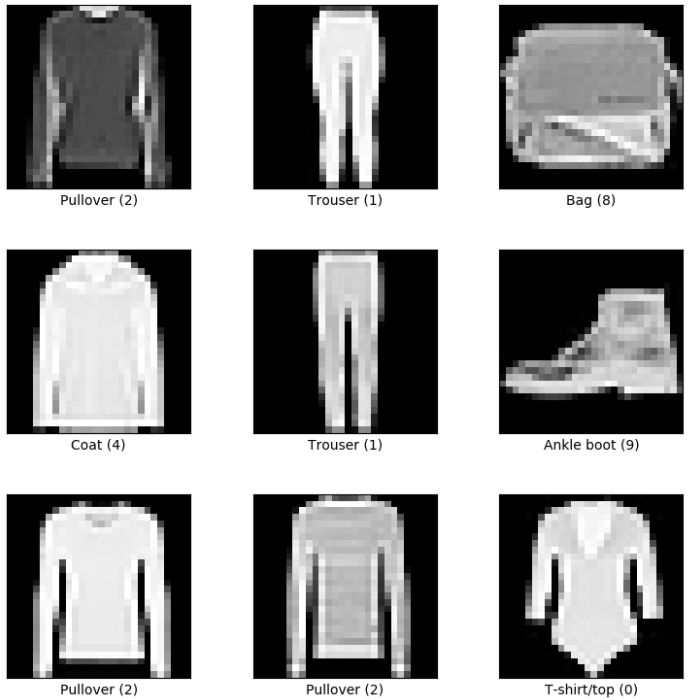
Traditional clustering
K-means, HDBSCAN, ...

Two-stages clustering
(PCA, t-SNE, UMAP, CAE) + (K-means, HDBSCAN)

Deep clustering
DCEC, ...

Methodology

- Analysis applied to images
- Three datasets
- Structure separated in three blocks
- Evaluation of the results
- Software



Traditional clustering

Do no use dimensionality reduction:
The clustering algorithm is good enough.

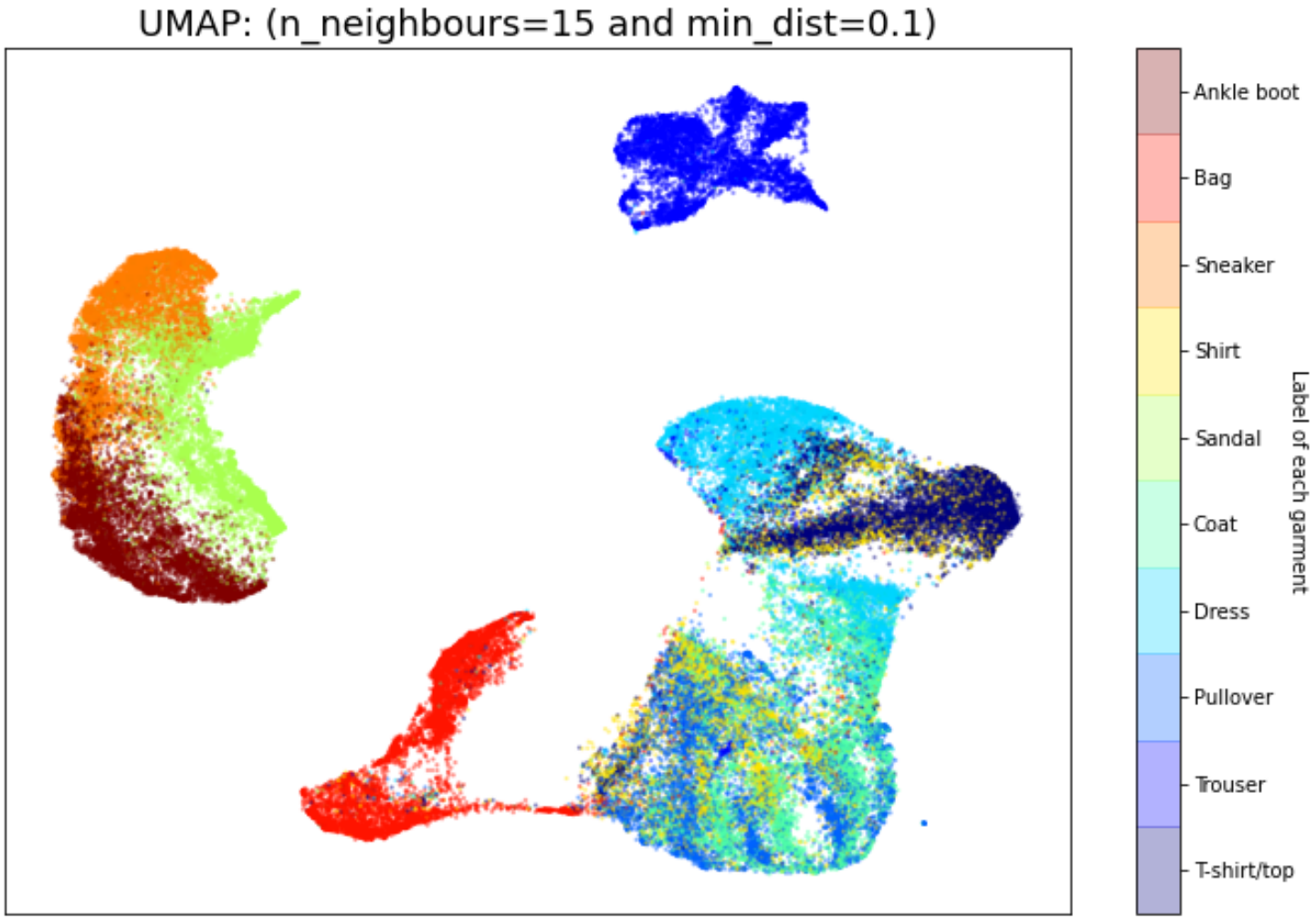
- Data have noise
- *The curse of dimensionality reduction*
- Results are limited by the algorithm

Two-stages clustering

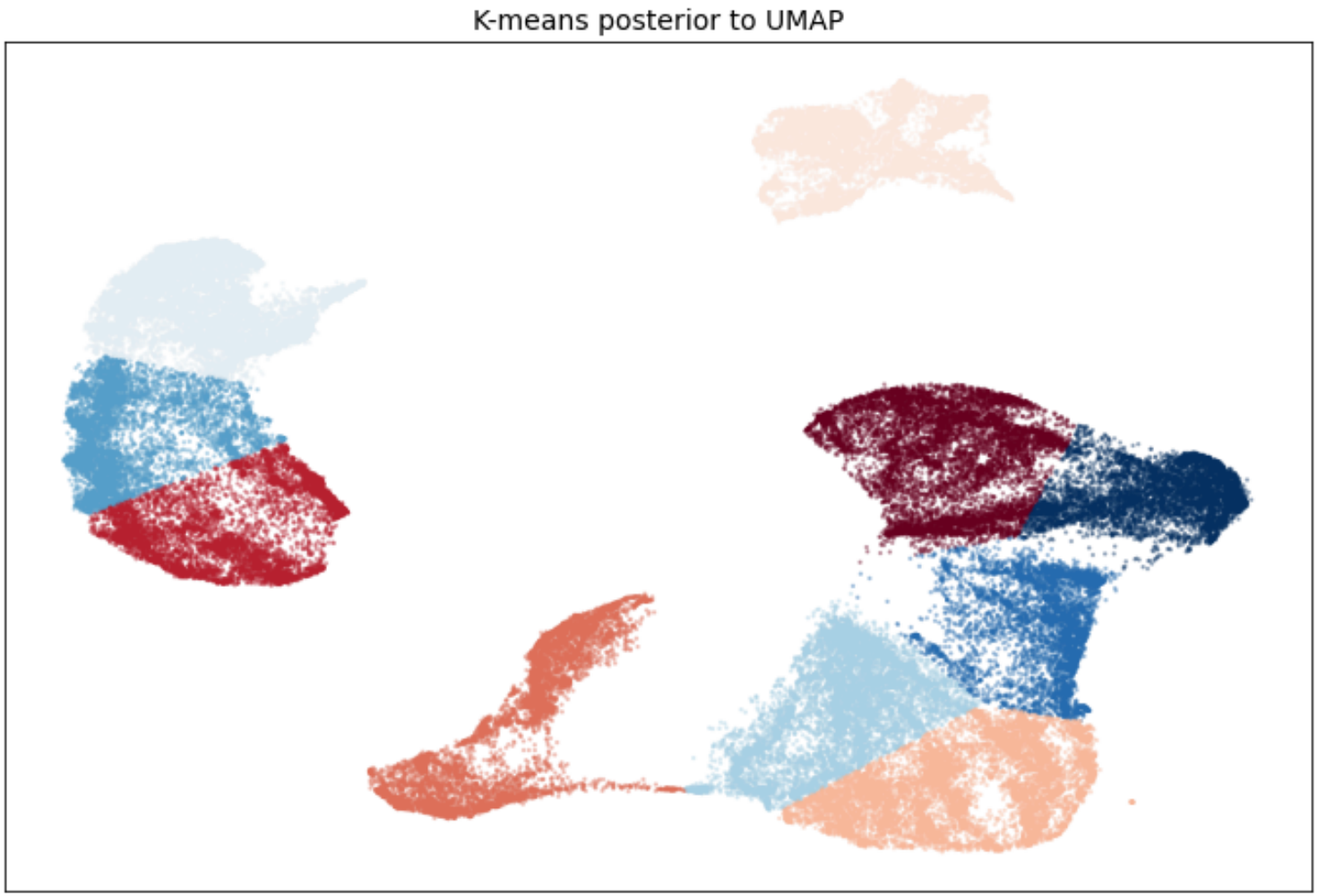
Apply clustering after DR:
A lower dimensional space boosts the performance of clustering

- Advantages
- Problems
- DEBATE: No clear conclusions.

First stage: DR



Second stage: Clustering



Results

Table: Quantitative performance of both strategies

Algorithm	NMI	ARI	ACC	SIL
K-means	0.5284	0.3844	0.5514	0.1517
PCA+HDBSCAN*	0.6621	0.4583	0.5275	-
t-SNE2+K-means	0.5842	0.4502	**0.6163**	-
UMAP2+K-means	**0.6433**	**0.4800**	0.5746	0.132
UMAP30+K-means	0.6373	0.4780	0.5833	0.088
UMAP2+HDBSCAN	0.6422	0.3970	0.4361	-
UMAP30+HDBSCAN	0.6064	0.2839	0.3450	-
CAE10+K-means	0.4862	0.3209	0.5087	0.0988

Deep clustering

Perform clustering and DR at the same time:

1. Good representations of the data are good for clutering
2. Clustering results provide supervised signal that improve the learning of representations

Advantages of neural networks

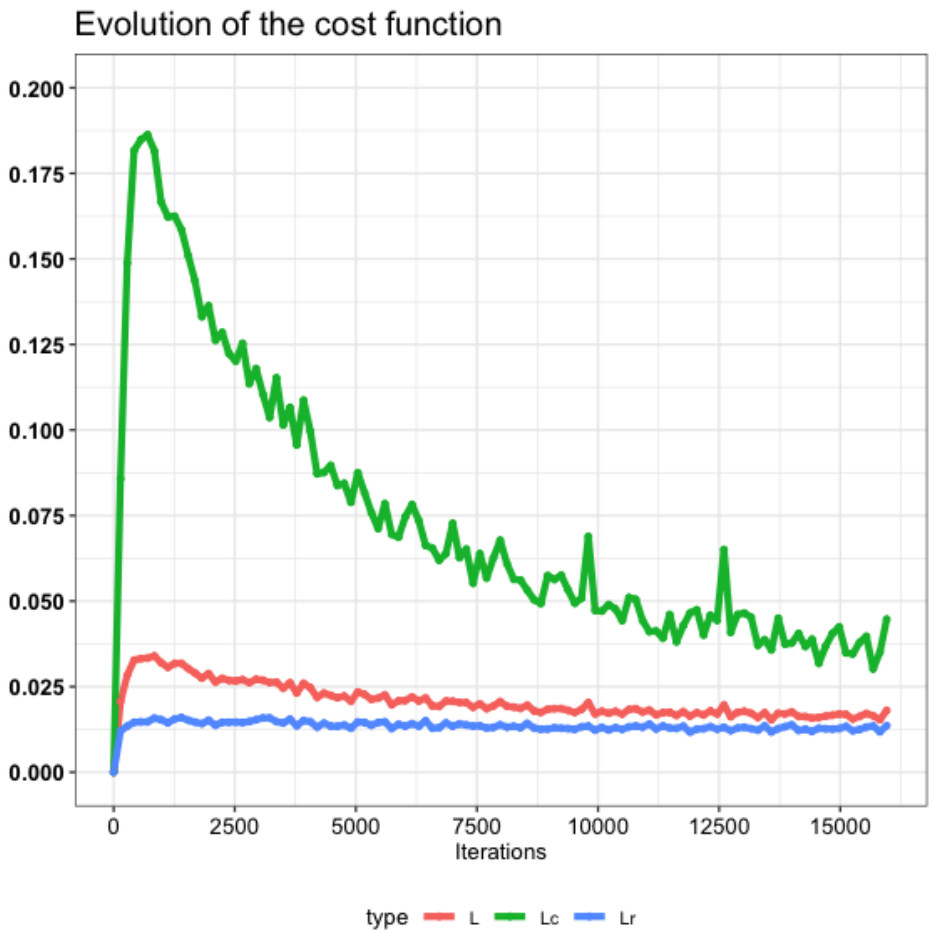
Deep convolutional embedding clustering (DCEC)

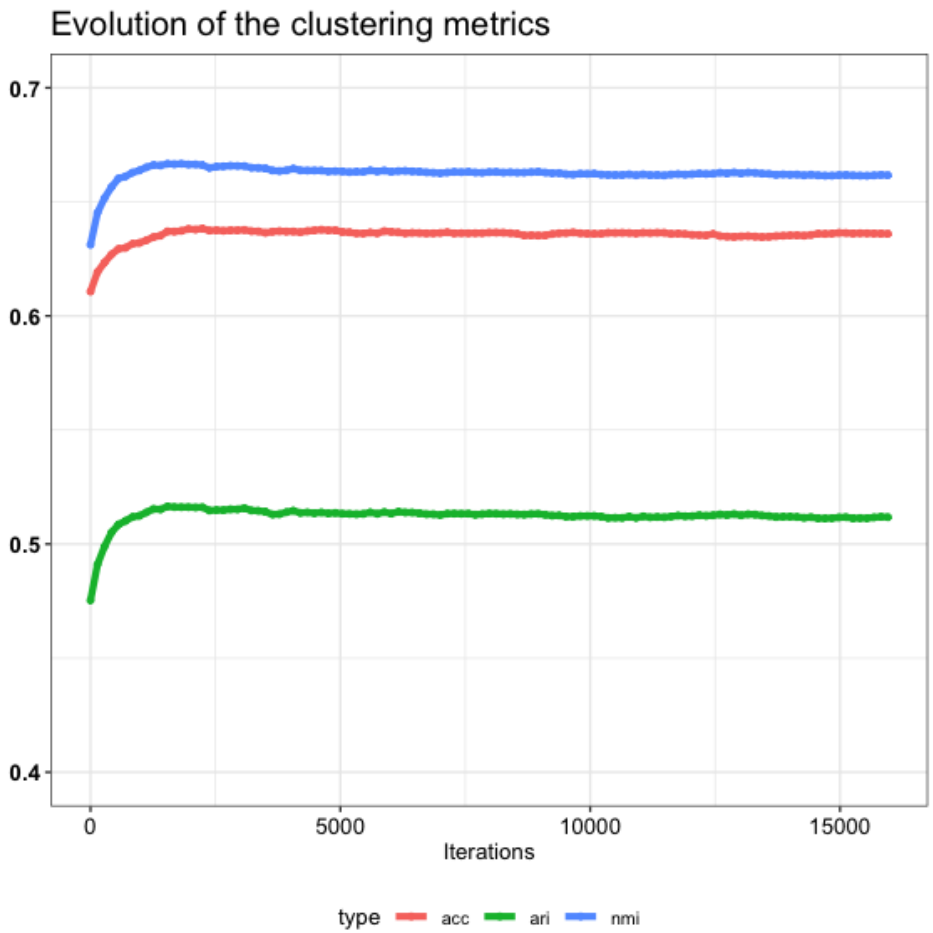
1. Pretraining
2. Parameter initialisation
3. Training

$$\text{CAE} \quad + \quad \text{K-means} + \text{t-SNE} \rightarrow \textbf{DCEC}$$

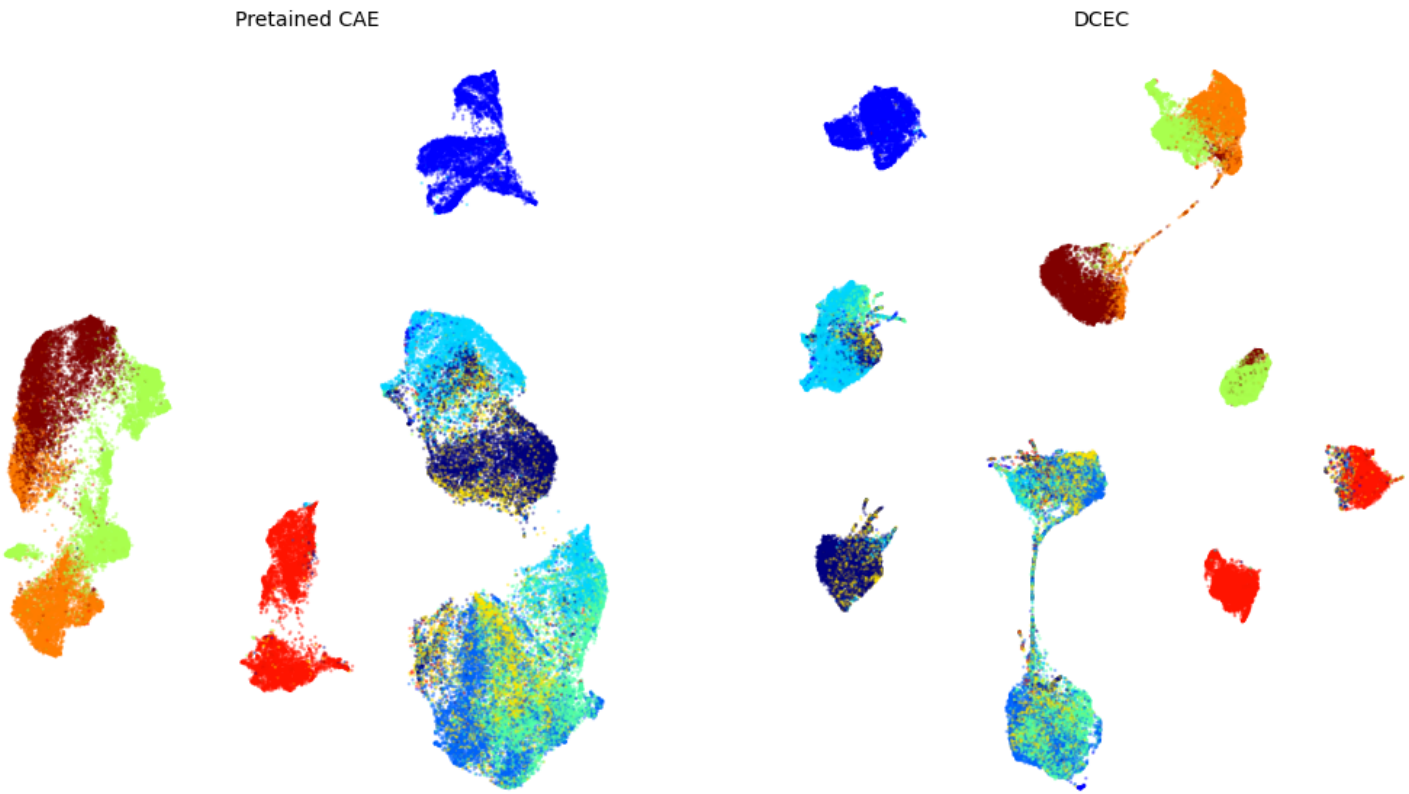
$$L = L_{rec} + \lambda L_{clust}$$

Training


Evolution of the cost function
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
2500
5000
7500
10000
12500
15000
Iterations
type
L
Lc
Lr


Evolution of the clustering metrics
0.7
0.6
0.5
0.4
0
5000
10000
15000
Iterations
type
acc
ari
nmi

Process



Final comparison

Table: Quantitative performance of Deep clustering

	Fashion			MNIST		
Algorithm	NMI	ARI	ACC	NMI	ARI	ACC
K-means	0.5284	0.3844	0.5514	0.4998	0.3665	0.5346
UMAP2+K-means	0.6433	0.4800	0.5746	0.8313	0.7451	0.8007
CAE10 Guo+K-means	0.5983	0.4308	0.5534	0.7882	0.7442	0.8365
CAE10+K-means	0.5430	0.3826	0.5149	0.5918	0.4801	0.615
DCEC Guo ($\gamma = 1$)	0.6072	0.4471	0.5609	0.5195	0.3792	0.5417
DCEC Guo	**0.6616**	**0.5117**	**0.6360**	**0.8898**	**0.8552**	**0.8924**

Conclusions

- Limitations of the analysis
- Large number of algorithms, both for DR and for clustering
- Additional applications of DR
- Progress of the field