Clustering validation by distribution hypothesis learning - Additional file

Ariel E. Bayá*[1] and Mónica G. Larese†[1]

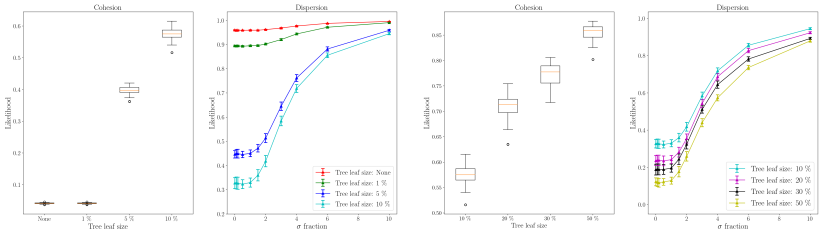
[1]CIFASIS, French Argentine International Center for Information and Systems Sciences, CONICET-UNR (Argentina) , Bv. 27 de Febrero 210 Bis, 2000 Rosario, Argentina

1 Additional material

References

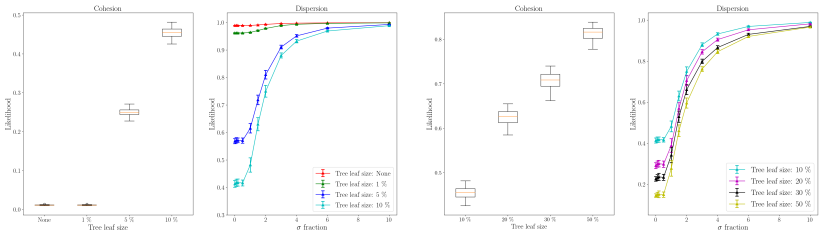
F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

*correspondece to: baya@cifasis-conicet.gov.ar
†larese@cifasis-conicet.gov.ar



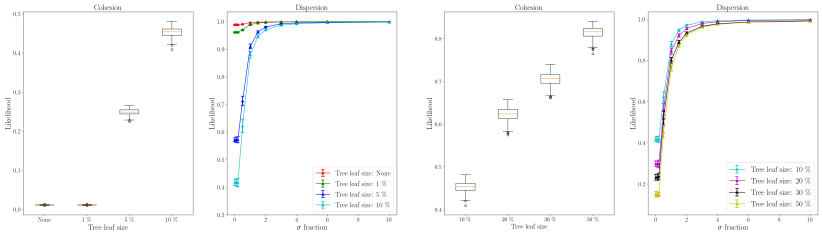
(a) Dimensionality 2

(b) Dimensionality 2



(c) Dimensionality 10

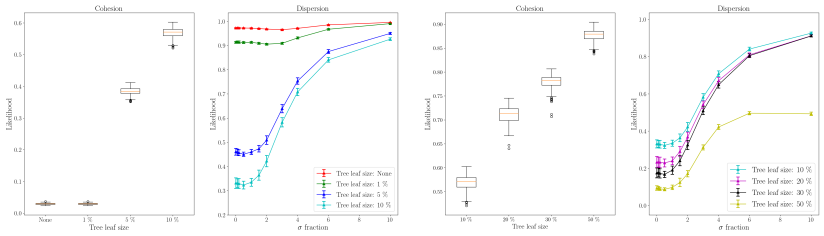
(d) Dimensionality 10



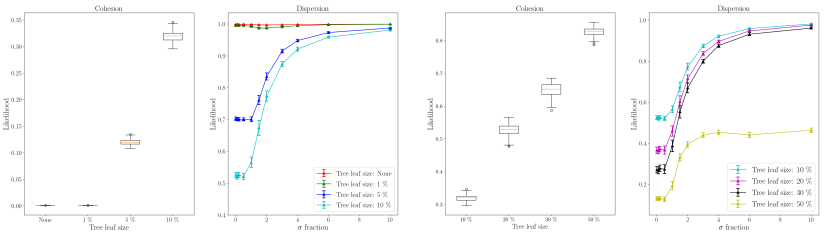
(e) Dimensionality 100

(f) Dimensionality 100

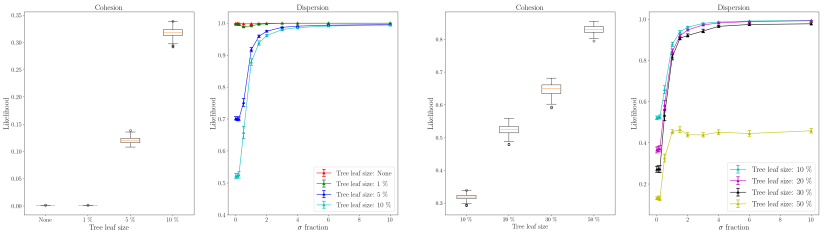
Figure 1: Cohesion-dispersion outcome for CART. We show two pairs of Cohesion-dispersion graphs. First, on the left, we display cohesion, and then, on the right, dispersion. Cohesion graphs show likelihood in the y-axis and the regularization parameter on the x-axis. Dispersion graphs are next to the right, and they present likelihood in the y-axis and cluster separation on the x-axis. There is one dispersion curve per regularization parameter. The figure rows show the same problem in increasing dimensions: 2, 10, and 100.



(a) Dimensionality 2 (b) Dimensionality 2

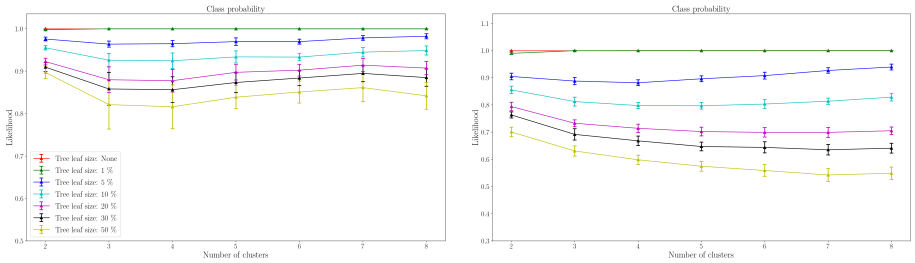


(c) Dimensionality 10 (d) Dimensionality 10



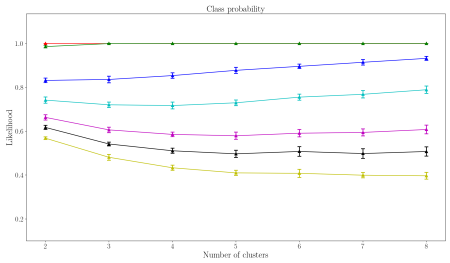
(e) Dimensionality 100 (f) Dimensionality 100

Figure 2: Cohesion-dispersion outcome for RF. We show two pairs of Cohesion-dispersion graphs. First, on the left, we display cohesion, and then, on the right, dispersion. Cohesion graphs show likelihood in the y-axis and the regularization parameter on the x-axis. Dispersion graphs are next to the right, and they present likelihood in the y-axis and cluster separation on the x-axis. There is one dispersion curve per regularization parameter. The figure rows show the same problem in increasing dimensions: 2, 10, and 100.



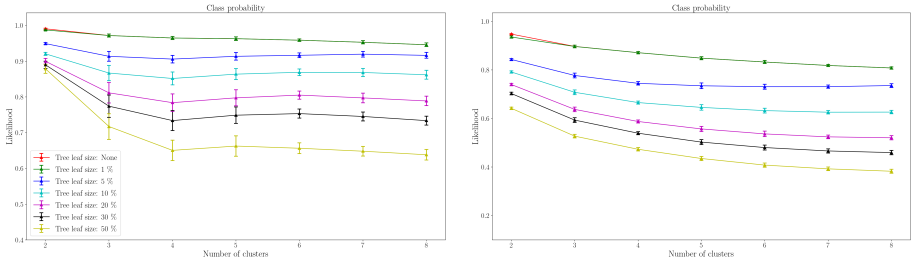
(a) Dimensionality 2

(b) Dimensionality 10



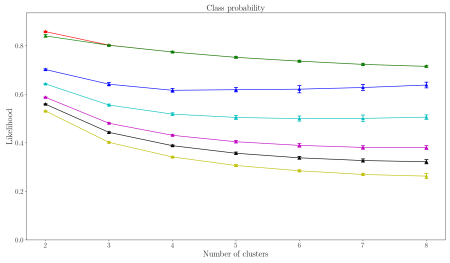
(c) Dimensionality 100

Figure 3: Solution likelihood outcome for CART. The curves show $E_X(p(Y_j = 1/X))$ for a varying number of clusters and regularization values. The data used is the same as in Figure 1. The dataset is analyzed for dimensionality (a) 2, (b) 10, and (100).



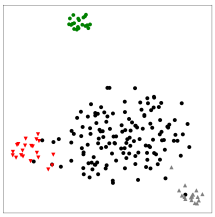
(a) Dimensionality 2

(b) Dimensionality 10

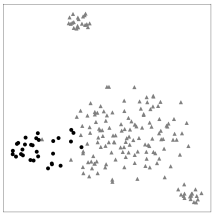


(c) Dimensionality 100

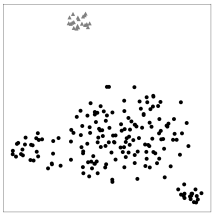
Figure 4: Solution likelihood outcome for RF. The curves show $E_X(p(Y_j = 1/X))$ for a varying number of clusters and regularization values. The data used is the same as in Figure 2. The dataset is analyzed for dimensionality (a) 2, (b) 10, and (100).



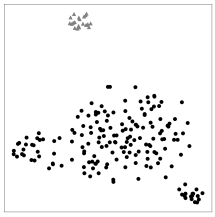
(a) Lung-A True



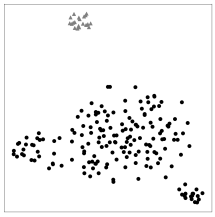
(b) Lung-A 2 clusters GMM



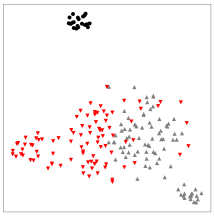
(c) Lung-A 2 clusters K-Means



(d) Lung-A 2 clusters HC-av

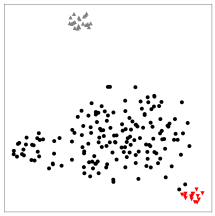


(e) Lung-A 2 clusters HC-co

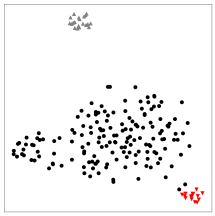


(f) Lung-A 3 clusters GMM

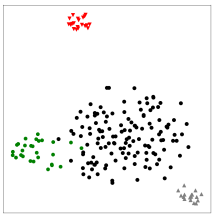
(g) Lung-A 3 clusters K-Means



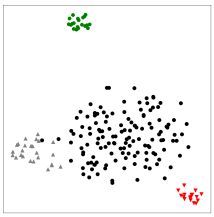
(h) Lung-A 3 clusters HC-av



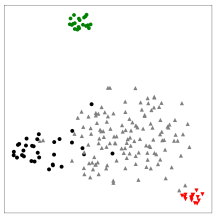
(i) Lung-A 3 clusters HC-co



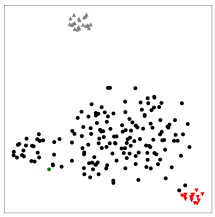
(j) Lung-A 4 clusters GMM



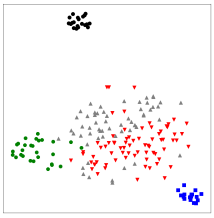
(k) Lung-A 4 clusters K-Means



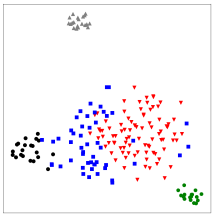
(l) Lung-A 4 clusters HC-av



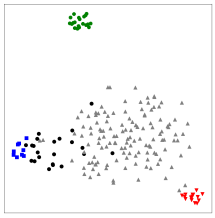
(m) Lung-A 4 clusters HC-co



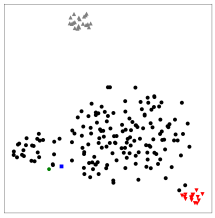
(n) Lung-A 5 clusters GMM



(o) Lung-A 5 clusters K-Means



(p) Lung-A 5 clusters HC-av



(q) Lung-A 5 clusters HC-co

Figure 5: Lung-A dataset. Panel (a) shows the true clusters scatter plot produced using Pedregosa et al. [2011] T-SNE version with the default setup. Then, clustering solutions are sorted by column according to $\{GMM, K-Means, HC-av, HC-co\}$. The figure rows show $\{2, 3, 4, 5\}$ clusters solutions.