Title: Tree-based Semi-supervised Clustering

Authors:

Claudio Conversano (University of Cagliari, ITALY),

Giulia Contu (University of Cagliari, ITALY),

Luca Frigau (University of Cagliari, ITALY)

Francesco Mola (University of Cagliari, ITALY)

ABSTRACT

Recently, in literature new clustering methods that take advantages of complex networks algorithms have been developed (De Oliveira et al., 2008; De Arruda et al., 2012). The use of complex networks approaches allows obtaining higher accuracy than traditional clustering methods. These new kinds of approaches consist in defining the distances between objects by classical metrics and performing complex network algorithms to partition the data, considering the distances as links for the derivation of the network (Granell et al., 2002; De Oliveira et al., 2008; Granell et al., 2010; De Arruda et al., 2012).

We propose a new approach in which the distances are defined by machine learning methods. Particularly, we consider tree-based methods such as Classification and Regression Tree (CART) and Random Forest (RF) (Breiman et al., 1984; Breiman, 2001). This approach results in a semi-supervised clustering method in which the communities are identified using community detection algorithms (traditional methods, divisive methods, modularity methods and other methods).

We illustrate the advantages of our approach through a simulation experiment and some real data examples. One of them concerns the performance metrics and the content characteristics of 137 websites of UNESCO sites located in Italy, France and Spain.

References

Breiman L. (2001), Random forests, Machine learning, vol. 45(1), pp. 5- 32.

Breiman L., Friedman J.H., Olshen R.A., Stone C.J. (1984), Classification Regression Trees, Wadsworth International Group, Belmont, California.

De Arruda G. F., Da Fontoura Costa L., Rodrigues F. A. (2012), A complex networks approach for data clustering, Physica A, 391, pp. 6174–6183.

De Oliveira T., Zhao L., Faceli K., de Carvalho A. (2008), Data clustering based on complex network community detection, in: IEEE Congress on Evolutionary Computation, 2008, CEC 2008 (IEEE World Congress on Computational Intelligence), pp. 2121–2126.

Granell C., Gomez S., Arenas A. (2012), Unsupervised clustering analysis: a multiscale complex networks approach, International Journal of Bifurction and Chaos, International Journal of Bifurcation and Chaos, vol. 22, available at https://arxiv.org/pdf/1101.1890.pdf.

Granell C., Gomez S., Arenas A. (2010), Data clustering using community detection, International Journal of Complex Systems in Science, vol. 1, pp. 21–24.