Proceeding Book of the
y-BIS Conference 2019:

Recent Advances in Data Science and Business Analytics

September 25 – 28, 2019
Mimar Sinan Fine Arts University / Fındıklı Campus

Edited by

Ozan Kocadagli
Ali Erkoc
Bilge Baser
Nihan Acar Denizli
Tahir Ekin
LOC of y-BIS 2019

Web design by
Ali Zafer Dalar

Cover photo by
Aylin Yaman Kocadagli

Cover design by
Ali Mertcan Kose

Damla Ilter

Published by

Mimar Sinan Fine Arts University Publications : 884

ISBN:

978-605-5005-95-5

Serial Number:

eMSGSÜ-FEF-İST-019/09-Kat1

Network-based Semisupervised Clustering

Giulia Contu*[1] , Luca Frigau[1], Claudio Conversano[1]

[1] Department of Economics and Business Sciences

Semisupervised learning [1] recently emerged as a new challenge: it uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task. In this framework, semi-supervised clustering ([2], chap. 20) concerns the application of clustering methods to partially labeled data or to data with other types of outcome measures. Most of the approaches to semisupervised clustering focused on the extension of standard clustering methods to the semisupervised setting. Some methods concern the situations when clusters are associated with a given outcome variable that acts as a "noisy surrogate". This variable is considered a good proxy of the unknown clustering structure if it is used in combination with the other variables in the case conventional clustering methods may fail in identifying suitable clusters. In this framework, the term "suitable" denotes consistency between the identified clusters and the levels of the outcome variable.

In this paper, we propose a novel approach to semi-supervised clustering that is inspired by the semisupervised clustering associated with an outcome variable approach mentioned above. The proposed approach combines an initialization step with a training step. In the first step, a classification or regression tree is grown considering the outcome variable as response and only one feature at a time and a specific weight, that depends on a variable importance measure arising with the single-feature tree, is assigned to each feature. At the same time, instances are weighted w.r.t. the homogeneity of the terminal node of the tree to which they are assigned. Next, single-feature trees are still learned in the training step. They use weighted instances and select randomly the feature used to grow the "single-feature tree" on the basis of the weights assigned to each feature in the previous iterations. The output of the tree allows to update a proximity matrix that counts how many times pairs of observations, say i and j, have been placed in the same terminal node of the tree. This matrix is the input for the derivation of a network N on which a community detection algorithm is applied. In each iteration, the community detection
algorithm provides a partition of the original set of n instances into k groups. The final (selected) partition is the one characterized by the minimum internal heterogeneity inside groups, the latter evaluated w.r.t. the outcome variable.

Keywords: Tree-based Classifiers; Complex Networks; Community Detection

References

[1] Olivier Chapelle, Alexander Zien and Bernhard Schölkopf (Eds.) Semi-supervised learning MIT Press, 2006.

[2] Charu C. Aggarwal and Chandan K. Reddy Data Clustering: algorithms and applications Chapman & Hall/CRC Press, 2014.

*Corresponding author: giulia.contu@unica.it