

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

<https://www.fupress.com/isbn/9788855183406>

ISSN 2704-601X (print)

ISSN 2704-5846 (online)

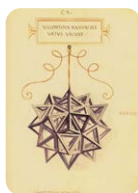
ISBN 978-88-5518-340-6 (PDF)

ISBN 978-88-5518-341-3 (XML)

DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs

Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard



Classification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

Firenze University Press Editorial Board

M. Garzaniti (Editor-in-Chief), M.E. Alberti, F. Vittorio Arrigoni, E. Castellani, F. Ciampi, D. D'Andrea, A. Dolfi, R. Ferrise, A. Lambertini, R. Lanfredini, D. Lippi, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, A. Orlandi, I. Palchetti, A. Perulli, G. Pratesi, S. Scaramuzzi, I. Stolzi.

📄 The online digital edition is published in Open Access on www.fupress.com.

Content license: except where otherwise noted, the present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: <http://creativecommons.org/licenses/by/4.0/legalcode>). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.

Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>).

© 2021 Author(s)

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

*This book is printed on acid-free paper
Printed in Italy*

NETWORK-BASED SEMI-SUPERVISED CLUSTERING OF TIME SERIES DATA

Claudio Conversano¹, Giulia Contu¹, Luca Frigau¹ and Carmela Cappelli²

¹ Department of Economics and Business, University of Cagliari, (e-mail: conversa@unica.it, giulia.contu@unica.it, frigau@unica.it)

² Department of Humanities, University of Naples Federico II, (e-mail: carmela.cappelli@unina.it)

ABSTRACT: Semisupervised clustering extends standard clustering methods to the semisupervised setting, in some cases considering situations when clusters are associated with a given outcome variable that acts as a “noisy surrogate”, that is a good proxy of the unknown clustering structure. A novel approach to semisupervised clustering associated with an outcome variable named network-based semisupervised clustering (NeSSC) has been recently introduced (Frigau *et al.*, 2021). It combines an initialization, a training and an agglomeration phase. In the initialization and training a matrix of pairwise affinity of the instances is estimated by a classifier. In the agglomeration phase the matrix of pairwise affinity is transformed into a complex network, in which a community detection algorithm searches the underlying community structure. Thus, a partition of the instances into clusters highly homogeneous in terms of the outcome is obtained. A particular specification of NeSSC, called Community Detection Trees (Co-De Tree), uses classification or regression trees as classifiers and the Louvain, Label propagation and Walktrap as possible community detection algorithm. NeSSC is based on an ad-hoc defined stopping criterion and a criterion for the choice of the optimal partition of the original data. In this presentation, we provide a new specification of the NeSSC algorithm that allows us to perform clustering of time series data. This specification is based on the integration between Co-De Tree and the Atheoretical Regression Tree (ART) approach introduced by (Cappelli *et al.*, 2013; Cappelli *et al.*, 2015). ART exploits the concept of contiguous partitions within the framework of Least Squares Regression Trees using as a single covariate an arbitrary sequence of completely ordered numbers $K = 1, 2, \dots, i, \dots, N$. Tree-regressing the response variable Y on this artificial covariate resorts to create and check at any node h all possible binary contiguous partitions of the $Y_i \in h$. These splits are the only ones that need to be checked to detect the binary partition that minimizes the sum of squares and, indeed, they are generated by using K as covariate. In other words, for the

contiguity property the best split lays in K (or in its subintervals after the split of the root node has taken place) and the tree algorithm, based on the classical “reduction in impurity” splitting criterion is forced to identify it. In general, the use of K as covariate enables ART to generate G different groups having different means. The effectiveness of the proposed NeSSC-ART combined approach for time series clustering is demonstrated on simulated and real data

KEYWORDS: network-based semisupervised clustering, community detection trees, atheoretical regression tree.

References

- CAPPELLI, C, D’URSO, P, & DI IORIO, F. 2013. Change point analysis of imprecise time series. *Fuzzi Sets and Systems*, **225**, 23–38.
- CAPPELLI, C, D’URSO, P, & DI IORIO, F. 2015. Regime change analysis of interval-valued time series with an application to PM10. *Chemometrics and Intelligent Laboratory System*, **146**, 337–346.
- FRIGAU, L, CONTU, G, MOLA, F, & CONVERSANO, C. 2021. *NNetwork-based semisupervised clustering*. Vol. 37.