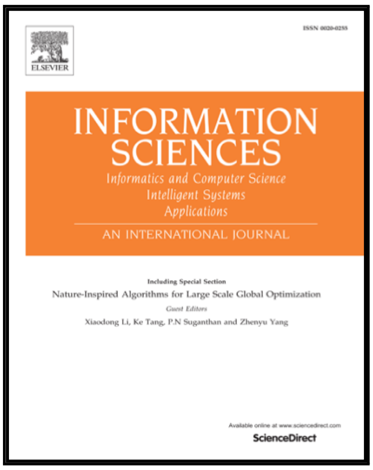

# Fuzzy clustering of distributional data with automatic weighting of variable components

Antonio Irpino, Rosanna Verde, Francisco de A.T. De Carvalho

PII: S0020-0255(17)30699-0
DOI: 10.1016/j.ins.2017.04.040
Reference: INS 12861

To appear in: *Information Sciences*

Received date: 9 May 2016
Revised date: 30 November 2016
Accepted date: 22 April 2017

# Fuzzy clustering of distributional data with automatic weighting of variable components

Antonio Irpino and Rosanna Verde[1]

*Universitá degli Studi della Campania "Luigi Vanvitelli", Dept. of Mathematics and Physics, 81100 Caserta, Italy*

Francisco de A.T. De Carvalho[2]

*Centro de Informatica, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes s/n - Cidade Universitaria, CEP 50740-560, Recife-PE, Brazil*

## Abstract

Distributional data, expressed as realizations of distributional variables, are new types of data arising from several sources. In this paper, we present some new fuzzy c-means algorithms for data described by distributional variables. The algorithms use the $L_2$ Wasserstein distance between distributions as dissimilarity measure. Usually, in fuzzy c-means, all the variables are considered equally important in the clustering task. However, some variables could be more or less important or even irrelevant for this task. Considering a decomposition of the squared $L_2$ Wasserstein distance, and using the notion of adaptive distance, we propose some algorithms for automatically computing relevance weights associated with variables, as well as with their components. This is done for the whole dataset or cluster-wise. Relevance weights express the importance of each variable, or of each component, in the clustering process acting also as a variable selection method. Using artificial and real-world data, we observed that algorithms with automatic weighting of variables (or components) are better able to take into account the cluster structure of data.

*Keywords:* Distribution-valued data, Wasserstein distance, Fuzzy clustering, Relevance weights, Adaptive distances
*2010 MSC:* 62H30, 62H86, 62A86

## 1. Introduction

One of the current big-data age requirements is the need of representing groups of data by summaries allowing the minimum loss of information as pos-

[1] antonio.irpino@unina2.it, rosanna.verde@unina2.it
[2] Corresponding Author: fatc@cin.ufpe.br