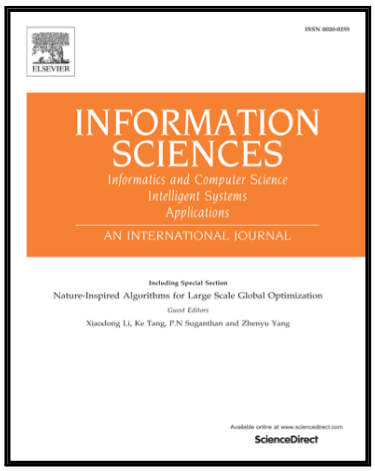

An Information-Theoretic Approach to Hierarchical Clustering of Uncertain Data

Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Sergio Greco

PII: S0020-0255(17)30626-6
DOI: 10.1016/j.ins.2017.03.030
Reference: INS 12813

To appear in: *Information Sciences*

Received date: 14 January 2015
Revised date: 6 February 2017
Accepted date: 25 March 2017

Please cite this article as: Francesco Gullo, Giovanni Ponti, Andrea Tagarelli, Sergio Greco, An Information-Theoretic Approach to Hierarchical Clustering of Uncertain Data, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.03.030

An Information-Theoretic Approach to Hierarchical Clustering of Uncertain Data

Francesco Gullo[a], Giovanni Ponti[b], Andrea Tagarelli[c], Sergio Greco[c]

[a]*UniCredit, R&D Dept., Rome, Italy*
[b]*ENEA, Portici Research Center, Portici (NA), Italy*
[c]*DIMES, University of Calabria, Rende (CS), Italy*

Abstract

Uncertain data clustering has become central in mining data whose observed representation is naturally affected by imprecision, staling, or randomness that is implicit when storing this data from real-word sources. Most existing methods for uncertain data clustering follow a partitional or a density-based clustering approach, whereas little research has been devoted to the hierarchical clustering paradigm. In this work, we push forward research in hierarchical clustering of uncertain data by introducing a well-founded solution to the problem via an information-theoretic approach, following the initial idea described in our earlier work [26]. We propose a prototype-based agglomerative hierarchical clustering method, dubbed *U-AHC*, which employs a new uncertain linkage criterion for cluster merging. This criterion enables the comparison of (sets of) uncertain objects based on information-theoretic as well as expected-distance measures. To assess our proposal, we have conducted a comparative evaluation with state-of-the-art algorithms for clustering uncertain objects, on both benchmark and real datasets. We also compare with two basic definitions of agglomerative hierarchical clustering that are treated as baseline methods in terms of accuracy and efficiency of the clustering results, respectively. Main experimental findings reveal that U-AHC generally outperforms competing methods in accuracy and, from an efficiency viewpoint, is comparable to the fastest baseline version of agglomerative hierarchical clustering.

Keywords: Clustering, Hierarchical clustering, Uncertain data, Information

Email addresses: gullof@acm.org (Francesco Gullo), giovanni.ponti@enea.it (Giovanni Ponti), tagarelli@dimes.unical.it (Andrea Tagarelli), greco@dimes.unical.it (Sergio Greco)