Pattern Recognition Letters 67 (2015) 122–129

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Clustering in augmented space of granular constraints: A study in knowledge-based clustering[☆]

Witold Pedrycz [a,*], Adam Gacek [b], Xianmin Wang [c]

[a] Department of Electrical & Computer Engineering, University of Alberta, Edmonton T6R 2V4 AB Canada & Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia and Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[b] Institute of Medical Technology and Equipment (ITAM), 118 Roosevelt Street, Zabrze 41-800, Poland
[c] Hubei Subsurface Multi-scale Imaging Key Laboratory, Institute of Geophysics and Geomatics, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Article history:
Available online 8 September 2015

Keywords:
Granular clustering
Information granules
Knowledge-based clustering
Granular constraints
Time series
Information granules of order-2 and type-2

ABSTRACT

In this study, a paradigm of fuzzy clustering is augmented by available domain knowledge expressed in the form of relational constraints built with the aid of a collection of fuzzy sets. These constraints are described as a collection of Cartesian products of fuzzy sets or their logic expressions are used to form an augmented data space and transform nonlinearly original data. Depending upon the nature of the constraints, discussed are two categories of resulting representations (clustering spaces), namely homogeneous spaces (in case when the transformations are fully expressed by means of the constraints) and heterogeneous spaces (when the resulting space is composed of some original variables present in the initially available data space and those being transformed and expressed by means of satisfaction levels of the constraints). The role of information granules of order-2 is revealed with regard to results of clustering produced in the transformed space. A generalization of the proposed approach is also discussed in case the clustered data are not numeric but are provided in the form of information granules; in this case a special attention is paid to a way in which a representation (description) of information granules is realized through relational constraints. We elaborate on the formation of the space (induced by constraints) and original data as well as discuss the detailed algorithmic developments.

1. Introductory notes

In a nutshell, clustering and fuzzy clustering, in particular, focus on discovering and describing structure in experimental data at a certain level of generality (abstraction) [30]. Clustering can be sought as a synonym of information granulation. Clusters are information granules. We bring together similar data that are significantly distinct from others so that they form some unique entities, which can be then processed and interpreted. Clustering is data-focused unsupervised processing. In contrast, knowledge-based clustering [17] is about realizing clustering processes being guided and augmented by available sources of knowledge usually articulated at a more abstract level than the data being clustered. Here we witness a plethora of existing schemes including conditional (context-based) clustering, proximity-based clustering, and collaborative clustering (Fig. 1)

There are two main directions of pursuits worth highlighting here; refer also to an overall taxonomy highlighting the essence of the arrangements of the algorithms:

(i) Clustering of numeric data with a focus on a follow up on interpretation of results and assessment of quality of clusters along with their applicability (e.g., in fuzzy modeling). Various formalisms of information granulation are sought leading to emergence of fuzzy clustering, rough clustering, shadowed set-based clustering and their various hybrid constructs [6,8,9];
(ii) Clustering information granules. Here the data themselves are information granules. There have been some early studies reported in [5,21].

In this sense, when alluding to granular clustering, there are two meanings of this term. First, we are concerned with the use of the technology of fuzzy sets, rough sets, etc. which leads to the formation of information granules expressed in the corresponding formalism [23]. Second, granular clustering is implied by the fact that the objects (patterns) being clustered are themselves information granules, cf. [14]. From the perspective of Granular Computing [14,29] and symbolic data analysis [3], in the first stream of investigations (i), we are interested in the design of information granules using various formalisms and their characterization of quality and discussing interpretability capabilities of the data on a basis of which they have been formed. Here a notion of reconstruction capabilities of

☆ This paper has been recommended for acceptance by Sankar Kumar Pal.

* Corresponding author. Tel.: +7804398731.
E-mail address: wpedrycz@ualberta.ca (W. Pedrycz).

http://dx.doi.org/10.1016/j.patrec.2015.08.019
0167-8655/

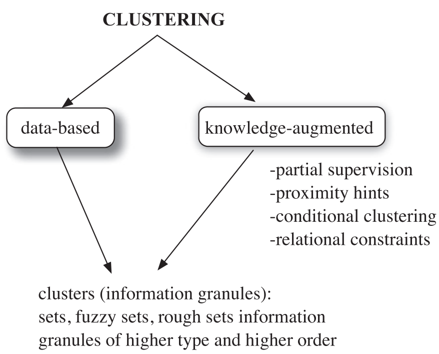

Fig. 1. A general taxonomy of data-driven and knowledge-oriented clustering.

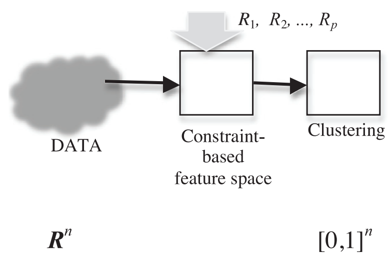

Fig. 2. A conceptual framework of relational constraint-based fuzzy clustering.

information granules is of interest [16,18]. There are also studies devoted to the description of data structure not in terms of numeric prototypes but rather granular prototypes with an intuitive underpinnings that representatives of diversified data need to be described by constructs at the higher level of abstraction, namely granular prototypes. In this way, we can witness a hierarchical form of processing in which processing at the higher level of the hierarchy involves information granules of higher type or higher order. This facet becomes also quite vividly present when discussing the second line of research (ii). Clustering information granules, see [14,24] is more challenging than dealing with numeric data and as a consequence, the produced structure of the data is captured by information granules of higher type.

The objective of this study is to propose an approach to clustering carried out in the space induced by information granules coming through a collection of relational (linguistic) constraints and study their properties. As the constraints can be viewed as a collection of knowledge tidbits, the study presented here falls under the umbrella of knowledge-based clustering. The essence of the problem can be succinctly outlined as follows. Given a collection of numeric data $\boldsymbol{X}$, reveal a structure of the data by taking into consideration a collection of relational constraints expressed by means of information granules and defined over all or selected variables of the original data space. Conceptually, two main phases are envisioned here. First, one has to accommodate the relational constraints, namely, they have to be formally described as a collection of information granules. Second, to cluster the data one has to directly engage these constraints. Here we propose a construction of a new data space, which directly reflects the constraints by and is directly implied (induced) by them. In this way, the original data are transformed to this new pace in which all clustering activities are completed. The essence of these two main development phases of the knowledge-based clustering is portrayed in Fig. 2.

The origin of relational constraints can be motivated by alluding to the following scenario that can be encountered when processing series of linked data sets (say, those collected over time such as encountered in data streams) and formed in the same feature space. Fuzzy clustering realized for the first of them, $\boldsymbol{D}_1$, gives rise to a collection of information granules. The obtained information granules are carefully analyzed and endowed with some interpretation. Denote the corresponding information granules (clusters) by $A_1, A_2, \ldots,$ and A_c. These information granules can be treated as reference fuzzy sets (viz. relational constraints). With the use of them we express new data so in this way A_is constitute a new space in which the data become represented. In the second time interval, collected is a new data set $\boldsymbol{D}_2$, and a study of its structure is carried out in the space of referential information granules already produced for $\boldsymbol{D}_1$. In other words, the data in $\boldsymbol{D}_2$ are expressed with the aid of A_is, namely for any $\boldsymbol{x}$ in $\boldsymbol{D}_2$ we compute some degrees of activation (membership) $A_1(\boldsymbol{x}), A_2(\boldsymbol{x}), \ldots, A_c(\boldsymbol{x})$. As the feature space in which when moving from one time window to the next one does not change, the new data are fully represented in terms of A_is and clustered in the new space induced by the structure obtained in the previous time interval. In other words the new structure (clusters) is interpreted with the use of A_is, which could be beneficial in the discovery of possible anomalies present in the data in the successive time window. Given the more abstract space in which clustering is completed now, a careful analysis of clustering results (prototypes) deserves attention. With regard it is worth mentioning an idea proposed in [27,28]. It is concerned with the use of filtering defined by Boolean constraints over Cartesian products of information systems in searching for relevant types of objects on higher levels of hierarchy.

In light of the formulation of the problem, the study exhibits several aspects of originality. Clustering in the space induced by relational constraints has not been investigated so far. The ways of forming new feature spaces come with some originality. The interpretability of the clustering results in the abstract space of constraints has not been studied. A general view at information granules advocated in this study, especially those emerging at the higher level of hierarchy (such as order-2 and type-2 information granules) is also insightful.

The paper is structured into six sections and follows a top-down exposure of the material. To make the material self-contained, Section 2 covers the fundamentals of Granular Computing. In Section 3, we briefly recall Fuzzy C-Means, FCM [2] in the context of this study. In Section 4, we focus on the formation of the feature space implied by the relational constraints; we also bring forward taxonomy of such spaces distinguishing between homogeneous and heterogeneous spaces. In the sequel, the results of clustering are supplied with a thorough interpretation where we highlight an increasing level of abstraction of the prototypes pointing at their linkages to order-2 information granules (Section 5). Section 6 is devoted to clustering of information granules; here an important facet of matching information granules and its quantification through possibility and necessity measures is underlined.

Throughout the study, we adhere to the standard notation encountered in pattern recognition and clustering. In what follows, we are concerned with a collection of N data, say $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$ located in $\boldsymbol{R}^n$. We also refer to some generic notions used in fuzzy sets. To focus our considerations, we concentrate on information granules formalized in terms of intervals and fuzzy sets. Likewise we refer to the FCM clustering scheme as a clustering vehicle. One has to note, however, that the main ideas and the proposed architecture of the clustering schemes apply equally well to other formalisms of information granules such as e.g., rough sets [13], shadowed sets [19], interval granules [10] etc. and algorithms of objective function-based clustering.

2. Selected prerequisites: granular computing

To make the study presented here self-contained and offer a better focus of the paper, we present a concise introduction to Granular Computing regarded as a formal vehicle to cast data analysis and modeling tasks in a certain conceptual framework.

Information granules are intuitively appealing constructs, which play a pivotal role in human cognitive and decision-making activities. We perceive complex phenomena by organizing existing knowledge