Knowledge-Based Systems 80 (2015) 98–108

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

# Data description: A general framework of information granules

Witold Pedrycz [a,b,c,*], Giancarlo Succi [d], Alberto Sillitti [d], Joana Iljazi [d]

[a] Department of Electrical & Computer Engineering, University of Alberta, Edmonton, T6R 2V4 AB, Canada
[b] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[c] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[d] Department of Computer Science, University of Bozen, I-39100 Bozen, Italy

ARTICLE INFO

*Article history:*
Received 22 September 2014
Received in revised form 17 December 2014
Accepted 30 December 2014
Available online 23 January 2015

*Keywords:*
Data description
Granular computing
Information granules
Principle of justifiable granularity
Fuzzy clustering
Software data
Interpretation

ABSTRACT

The study is concerned with a granular data description in which we propose a characterization of numeric data by a collection of information granules so that the key structure of the data, their topology and essential relationships are described in the form of a family of fuzzy sets – information granules. A comprehensive design process is introduced in which we show a two-phase development strategy: first, numeric prototypes are built with the use of Fuzzy C-Means (FCM) that is followed by their augmentation resulting in a collection of information granules. In the design of information granules we engage the fundamental ideas of Granular Computing, especially the principle of justifiable granularity. A series of experiments is presented to visualize the key steps of the construction of information granules.



## 1. Introduction

Data description has been one of the key pursuits in the broad plethora of data analysis. The need for a concise, highly interpretable, and accurate descriptors of data is highly visible so that such descriptions reveal and describe an essence of the main relationships and associations among variables of the systems. We have been witnessing a slew of approaches originating from studies developed within the setting of statistical analysis [19]. Quite often data description is referred to as anomaly detection as we are predominantly concerned with a single-class problem where a class of interest (to be described) is the one for which we are to form a collection of descriptors [8]. A number of investigations and proposals arose within the realm [1,7,9]. There has been a direction to support an abstract view at data and their description such as the one arising in the realm of symbolic data analysis [3].

Having in mind the key objectives of data description where the concern is both on meaningfulness (relevance) of the descriptors of data and interpretability of these descriptors. The problem is exacerbated given a diversity of data. On the one hand, we encounter large data sets of high dimensionality. On the other, one has to deal with data exhibiting a very limited number of records but characterized by high dimensionality. The main point underlined here is that in order to address the important objectives of relevance and interpretability of data descriptors, the description mechanisms of data and the descriptors themselves arising therein have to become inherently information granules rather than numeric entities.

The formal framework of processing is based on Granular Computing [10,11,20,21], which is focused on developing, characterizing and processing information granules. Let us recall that information granules are regarded as abstract constructs that bring together elements of some closeness (resemblance) – in the context of the problem discussed here those are elements that describe and are representative of the collection of these elements to a significant extent.

Let us start with some qualitative setting and highlight the essence of the problem in a two-dimensional case. Consider a collection of data belonging to a given class (black dots) we would like to describe (characterize), see Fig. 1. There are also some data belonging to another class (shown by squares); their number is small and the data themselves are far more scattered and irregularly distributed across the space in several cases overlapping with the regions with the high density of data belonging to the class to be described.

One can capture the essence of the groups of the data visualized there by forming some geometric constructs embracing the data. Intuitively, these descriptors should include as many data points coming from the class we intend to describe while at the same time leaving out (excluding) the data not belonging to the class

* Corresponding author at: Department of Electrical & Computer Engineering, University of Alberta, Edmonton, T6R 2V4 AB, Canada.
*E-mail address:* wpedrycz@ualberta.ca (W. Pedrycz).

http://dx.doi.org/10.1016/j.knosys.2014.12.030
0950-7051/

of interest. The geometric descriptors could be highly diversified as shown in Fig. 1(b). They could be made more regular as those illustrated in Fig. 1(c) where the data are "covered" by a collection of rectangles. While the first option delivers a great deal of flexibility, one may anticipate that their construction could be more demanding and their compact interpretation might cause some difficulties. On the other hand, the rectangular shapes of descriptors come with an intuitively appealing interpretation as a Cartesian product of a collection of intervals formed over the individual variables, say $[a, b] \times [w, z]$; see Fig. 1(c). Evidently, the geometry we are dealing with now is simpler than in the one being captured by the sophisticated geometric figures shown in Fig. 1(b). It is also more interpretable. In the same vein, we can talk about a description realized by fuzzy sets or rough sets.

The ultimate objective of this study is to develop a granular description of data where the crux of the data and the dominant topology of the data are well represented. We establish a two-phase development process. While the first phase is based on the formation of the numeric structure of the data (captured through a series of numeric prototypes), the second phase offers a substantial enhancement of the description of the structure by forming information granules. The characterization of the granules in terms of their coverage, specificity as well as their geometric localization bring about a detailed insight into the data's description.

While there have been a number of studies devoted to a single class data description and classification, the originality of the investigations reported in this paper is at least twofold. First, the proposed approach is general as we engage a comprehensive environment of Granular Computing forming a conceptual framework of data description. Second, we provide a comprehensive algorithmic scheme showing on how information granules in the interval form can be constructed. Here we form information granules following the principle of justifiable granularity here a sound balance between specificity and experimental justifiability of the granules is achieved.

The study is structured as follows. We cover some fundamentals of Granular Computing, which build all required prerequisites and help cast the study in a general setting (Section 2). In Section 3, we provide a formal formulation of the problem while in Section 4 briefly recall Fuzzy C-Means as a generic mechanism to cluster data and build fuzzy clusters. The buildup of granular prototypes realized on a basis of numeric prototypes produced by the FCM method is discussed in Section 5 where the principle of justifiable granularity is studied. An overall architecture of the process starting from numeric data and resulting in granular prototypes and their characterization is elaborated on in Section 6. Experimental studies are covered in Section 7.

## 2. Information granules and Granular Computing

To make the study presented here self-contained and offer a better focus, we present a concise introduction to Granular Computing regarded as a formal vehicle to cast data analysis 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 along with available experimental evidence and structuring them in a form of some meaningful, semantically sound entities, which are central to all ensuing processes of describing the world, reasoning about the environment and support decision-making activities. The term information granularity itself has emerged in different contexts and numerous areas of application. It carries various meanings. One can refer to Artificial Intelligence in which case information granularity is central to a way of problem solving through problem decomposition where various subtasks could be formed and solved individually. In general, by information granule one regards a collection of elements drawn together by their closeness (resemblance, proximity, functionality, etc.) articulated in terms of some useful spatial, temporal, or functional relationships. Subsequently, Granular Computing is about representing, constructing, and processing information granules.

We can refer here to some areas, which offer compelling evidence as to the nature of underlying processing and interpretation in which information granules play a pivotal role: image processing, processing and interpretation of time series, granulation of time, design of software systems.

Information granules are examples of abstractions. As such they naturally give rise to hierarchical structures: the same problem or system can be perceived at different levels of specificity (detail) depending on the complexity of the problem, available computing resources, and particular needs to be addressed. A hierarchy of information granules is inherently visible in processing of information granules. The level of detail (which is represented in terms of the size of information granules) becomes an essential facet facilitating a way a hierarchical processing of information with different levels of hierarchy indexed by the size of information granules.

Even such commonly encountered and simple examples presented above are convincing enough to lead us to ascertain that (a) information granules are the key components of knowledge representation and processing, (b) the level of granularity of information granules (their size, to be more descriptive) becomes crucial to the problem description and an overall strategy of problem solving, (c) hierarchy of information granules supports an important aspect of perception of phenomena and deliver a tangible way of dealing with complexity by focusing on the most essential facets of the problem, (d) there is no universal level of granularity of information; commonly the size of granules is problem-oriented and user dependent.

There are several well-known formal settings in which information granules can be expressed and processed:

*Sets (intervals)* realize a concept of abstraction by introducing a notion of dichotomy: we admit element to belong to a given information granule or to be excluded from it. Along with set theory

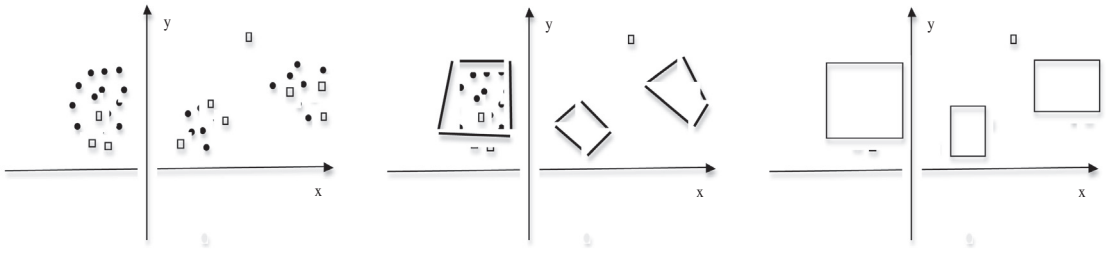


**Fig. 1.** Example data (black dots) to be described in a two-dimensional space; shown are also data belonging to the second class (squares) to be excluded from the descriptors formed for the first class (a) along with a collection of geometric descriptors of diverse shape (b) and of rectangular shape (c).