Pattern Recognition Letters 67 (2015) 113–121

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Data granulation by the principles of uncertainty ☆

Lorenzo Livi*, Alireza Sadeghian

Department of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada

ARTICLE INFO

Article history:
Available online 29 April 2015

Keywords:
Data granulation
Granular modeling and computing
Principles of uncertainty
Uncertainty measure
Type-2 fuzzy set

ABSTRACT

Researches in granular modeling produced a variety of mathematical models, such as intervals (higher-order) fuzzy sets, rough sets, and shadowed sets, which are all suitable to characterize the so-called information granules. Modeling of the input data uncertainty is recognized as a crucial aspect in information granulation. Moreover, the uncertainty is a well-studied concept in many mathematical settings, such as those of probability theory, fuzzy set theory, and possibility theory. This fact suggests that an appropriate quantification of the uncertainty expressed by the information granule model could be used to define an invariant property, to be exploited in practical situations of information granulation. In this perspective, we postulate that a procedure of information granulation is effective if the uncertainty conveyed by the synthesized information granule is in a monotonically increasing relation with the uncertainty of the input data. In this paper, we present a data granulation framework that elaborates over the principles of uncertainty introduced by Klir. Being the uncertainty a mesoscopic descriptor of systems and data, it is possible to apply such principles regardless of the input data type and the specific mathematical setting adopted for the information granules. The proposed framework is conceived (i) to offer a guideline for the synthesis of information granules and (ii) to build a groundwork to compare and quantitatively judge over different data granulation procedures. To provide a suitable case study, we introduce a new data granulation technique based on the minimum sum of distances, which is designed to generate type-2 fuzzy sets. The automatic membership function elicitation is completely based on the dissimilarity values of the input data, which makes this approach widely applicable. We analyze the procedure by performing different experiments on two distinct data types: feature vectors and labeled graphs. Results show that the uncertainty of the input data is suitably conveyed by the generated type-2 fuzzy set models.

1. Introduction

Granulation of information [11,16,38,42,43] emerges as an essential data analysis paradigm. Information used or acquired to describe an abstract/physical/social process is usually expressed in terms of data (experimental evidence). Therefore, granulation of information usually translates to data granulation. Granulation of data can be roughly described as the action of aggregating semantically and functionally similar elements of the available experimental evidence. This is performed to achieve a higher-level data description, which is implemented in terms of information granules (IGs) [32]. IGs are sound data aggregates that are formally described by a suitable mathematical model. Many mathematical settings have been proposed so far in the related literature, such as intervals—hyperboxes (higher order) fuzzy sets, rough sets, and shadowed sets [38]. The synthesized IGs can be used for interpretability purposes [26,27] or they can be used as a computational component of a suitable intelligent system [1,4,10,22,24,25,32,34,37,50]. Nonetheless, the problem of designing effective and justifiable data granulation procedures (GPs) remains of paramount importance [6,7,12,29–31,40,41].

The principle of justifiable granularity (PJG) is a well-established guideline for the synthesis of IGs [33,35,36]. The PJG states that granulation should be performed by finding the "optimal" compromise among two conflicting requirements: specificity and generality. In other terms, an IG modeling input data should be designed such that it retains only the essential information (it should be specific, conveying a specific semantic content) but, at the same time, it should cover a reasonable amount of information. Since the PJG is conceived to provide an adaptive mechanism to the information granulation problem, it is not designed to directly offer a built-in mechanism to objectively evaluate the quality of the granulation itself. To this end, it is necessary to rely on external performance measures to quantify and judge over the quality of an IG.

The uncertainty is a peculiar property of virtually every human action that involves reasoning, decision making, and perception [44,48].

☆ This paper has been recommended for acceptance by Gabriella Sanniti di Baja.

* Corresponding author. Tel.: +1 416 979 5000; fax: +1 416 979 5064.
E-mail address: llivi@scs.ryerson.ca, lorenz.livi@gmail.com (L. Livi).

http://dx.doi.org/10.1016/j.patrec.2015.04.008
0167-8655/

Modeling the uncertainty of the input data is an essential mission in data granulation. In fact, any IG model is designed to handle and hence express the uncertainty through an appropriate formalism. How the uncertainty is embedded into the IG model depends, of course, on the specific mathematical setting used for the IG. However, while the numerical quantification of the uncertainty pertaining a specific situation may change as we change the mathematical setting of the IG, the *level* of uncertainty should remain the same. In these terms, the principles of uncertainty [14,15] offer a compelling guideline to implement and evaluate practical data granulation techniques.

In this paper, we elaborate a conceptual data granulation framework over the principles of uncertainty. A preliminary version of the herein exposed ideas appeared in [20]. Here we further elaborate over this preliminary work by providing a more extensive discussion of the framework, offering new experiments that demonstrate the different facets underlying such ideas. In the proposed framework we idealize the uncertainty as an "invariant" property, to be preserved as much as possible during the granulation of the input data. As a consequence, we are able to objectively quantify the effectiveness of the granulation, regardless of the input data representation and the adopted IG model. This interpretation allows also to quantitatively judge on a common groundwork different data granulation techniques operating on the same data. We provide a demonstration of these ideas by discussing a data granulation technique that generates type-2 fuzzy sets (T2FSs).

This paper is structured as follows. Section 2 introduces the principles of uncertainty. Throughout Section 3 we introduce the proposed conceptual framework for data granulation. In Section 4 we present a procedure to generate T2FSs by means of the minimum sum of distances (MinSOD) technique. In Section 5 we discuss the experiments and related results. Section 6 concludes the paper. We provide two appendices: Appendix A introduces to the context of T2FSs, while Appendix B the MinSOD.

2. The principles of uncertainty

The principles of uncertainty have been introduced by [14] two decades ago, with the aim of providing high-level guidelines to the development of well-justified methods for problem solving in presence of uncertainty. Such principles elaborate over the ubiquitous concepts of uncertainty and information. It is intuitive to understand that uncertainty and information are intimately related: the reduction of uncertainty is caused by gaining new information, and vice versa.

Three principles have been introduced (quotes are taken from [14]):

1. Principle of minimum uncertainty: "It facilitates the selection of meaningful alternatives from solution sets obtained by solving problems in which some of the initial information is inevitably reduced in the solutions to various degrees. By this principle, we should accept only those solutions in a given solution set for which the information reduction is as small as possible."
2. Principle of maximum uncertainty: "This is reasoning in which conclusions are not entailed in the given premises. Using common sense, the principle may be expressed by the following requirement: in any ampliative inference, use all information available but make sure that no additional information is unwittingly added."
3. Principle of uncertainty invariance: "The principle requires that the amount of uncertainty (and information) be preserved when a representation of uncertainty in one mathematical theory is transformed into its counterpart in another theory."

A combination of the first and third principle provides a compelling guideline for the purpose of data granulation. In fact, granulation implies mapping some input data (experimental evidence) originating from a certain input domain, say $\mathcal{X}$, to a domain of IGs, say $\mathcal{Y}$. We argue that, when performing such a mapping, the uncertainty, regardless of the adopted formal mathematical framework, should be considered as an invariant property to be preserved as much as possible.

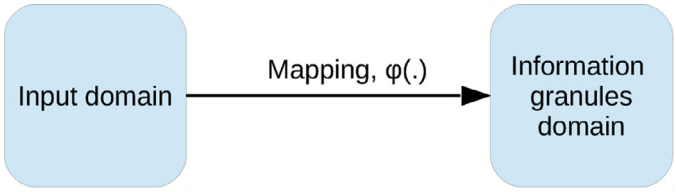

Fig. 1. Data granulation as a mapping, $\phi : \mathcal{P}_{<\infty}(\mathcal{X}) \to \mathcal{Y}$.

3. Data granulation with the principles of uncertainty

In this section, we introduce the proposed data granulation framework. Fig. 1 illustrates the data granulation process. A procedure of data granulation can be formalized as a mapping, $\phi(\cdot)$, among two domains: input domain, $\mathcal{X}$, and the output domain, $\mathcal{Y}$. $\mathcal{X}$ is the domain of the input data, whereas $\mathcal{Y}$ is a domain of IGs (e.g., a domain of hyperboxes, fuzzy sets, shadowed sets, rough sets and so on). In practice, $\phi(\cdot)$ is a formal procedure for mapping a finite input dataset $\mathcal{S} \in \mathcal{P}_{<\infty}(\mathcal{X})$ with an output IG, say $\widetilde{\mathcal{A}} \in \mathcal{Y}$, i.e., $\widetilde{\mathcal{A}} = \phi(\mathcal{S})$. Please note that we used a special mapping, $\mathcal{P}_{<\infty}(\cdot)$, in the input domain to allow discussing about $\mathcal{S}$ in terms of "element" of the input domain; in the following $\mathcal{P}_{<\infty}(\mathcal{X})$ is assumed to return all n-subsets of $\mathcal{X}$, with n finite. Note that $\mathcal{Y}$, as well as $\widetilde{\mathcal{A}}$, should be denoted by making explicit reference to $\mathcal{X}$ and $\mathcal{S}$, respectively, since IGs depend on the input. However, if no confusion is possible, we will avoid such specifications.

There are a number of important questions that should be answered: "Is the mapping $\phi(\cdot)$ well-justified? Moreover, how do we objectively assess the quality of the mapping?" "Are there invariant properties that must be preserved in the transformation from $\mathcal{S}$ to $\widetilde{\mathcal{A}}$?" "Can we numerically quantify those properties?" "Given two GPs, are we able to affirm that one performs a better granulation than the other by considering the same experimental conditions?" Reasoning over those questions provides important motivations for the design and formal evaluation of information GPs.

IGs are semantically sound constructs that are synthesized to convey higher-level information with respect to (w.r.t.) the data from which they are generated [32]. All models used in information granulation [38] are designed to realize a "simplification" of the input data. This consists in aggregating data that are considered indistinguishable (indiscernible) and functionally/semantically related. IGs are hence designed also to handle the uncertainty caused by this simplification. How the uncertainty is handled by the IG model depends on the specific mathematical setting used to describe the IG [15]. However, it is a reasonable assumption that, regardless of the specific mathematical setting, two IGs with different models, but synthesized from the same input data, should convey a comparable uncertainty, i.e., they should agree at least on the "level of uncertainty". The same concept holds for the uncertainty measured in the input with the one measured in the resulting output IG.

In the following, we formalize a conceptual framework to design and evaluate specific implementations of the mapping $\phi(\cdot)$. We refer to the proposed framework as the principle of uncertainty level preservation (PULP). Usually, $\mathcal{X}$ is a domain of non-granulated data, such as $\mathbb{R}^d$ vectors, sequences of objects, or graphs. However, $\mathcal{X}$ can be conceived also as a domain of IGs. In this case, since the role of $\phi(\cdot)$ is to provide an abstraction, $\mathcal{Y}$ must be a domain of higher-level IGs w.r.t. those of $\mathcal{X}$. In the following, however, we will consider mappings from input domains of non-granulated data types only.