Applied Soft Computing 46 (2016) 307–316

Contents lists available at ScienceDirect

Applied Soft Computing

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

Granular data imputation: A framework of Granular Computing

Chunfu Zhong[a], Witold Pedrycz[a,b,c,*], Dan Wang[a], Lina Li[a], Zhiwu Li[d,e]

[a] School of Electro-Mechanical Engineering, Xidian University, Xi'an 710071, PR China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2R3, Canada
[c] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[d] Institute of Systems Engineering, Macau University of Science and Technology, Taipa, Macau
[e] Faculty of Engineering, King Abdulaziz University, Jeddah 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 4 May 2015
Received in revised form 27 April 2016
Accepted 3 May 2016
Available online 19 May 2016

Keywords:
Data imputation
Granular Computing
Reconstruction
Granular data
Principle of justifiable granularity
Fuzzy clustering

ABSTRACT

Data imputation is a common practice encountered when dealing with incomplete data. Irrespectively of the existing spectrum of techniques, the results of imputation are commonly numeric meaning that once the data have been imputed they are not distinguishable from the original data being initially available prior to imputation. In this study, the crux of the proposed approach is to develop a way of representing imputed (missing) entries as information granules and in this manner quantify the quality of the imputation process and the quality of the ensuing data. We establish a two-stage imputation mechanism in which we start with any method of numeric imputation and then form a granular representative of missing value. In this sense, the approach could be regarded as an enhancement of the existing imputation techniques.

Proceeding with the detailed imputation schemes, we discuss two ways of imputation. In the first one, imputation is realized for individual variables of data sets and afterwards enhanced by the buildup of information granules. In the second approach, we are concerned with the use of fuzzy clustering, Fuzzy C-Means (FCM), which helps establish a structure in the data and then use this information in the imputation process.

The design of information granules invokes the fundamentals of Granular Computing, namely a principle of justifiable granularity and an allocation of information granularity. Numeric experiments concerned with a suite of publicly available data sets offer detailed insights into the main facets of the overall design process and deliver a parametric analysis of the methods.

1. Introductory notes

Imputation of data [4,12,22,26,28,29,31] is one of the important activities associated with enhancing data quality. It is often regarded as a prerequisite for any further processing (classification, prediction) in which the data are going to be used [13]. Missing items have to be prudently imputed, otherwise biased results may cause poor performance of the ensuing constructs [24]. The literature on this subject is very diversified, and the proposed methods of imputation vary in terms of their assumptions, sophistication and reported nature of the results [6,11,20,21,25,32]. These methods realize single imputation and multiple imputation [26]. Some of the methods falling under the first group include mean imputation, regression imputation [23], and hot deck imputation [1]. Multiple imputation techniques are reported in Refs. [24,25]. It is quite an agreeable position that there are no ideal methods and in many cases, as reported in the literature [7], some simple methods may perform equally well as more advanced techniques. The difficulty in the assessment of the efficiency of a specific imputation algorithm and reported results may vary from case to case.

Interestingly, there is a striking similarity among all the methods: once the imputation has been completed, the imputed data cannot be told apart from the original data, which have not been affected through the imputation process. Intuitively, it could have been expected that the imputed data should manifest in a different way than the originally available numeric data. In the sequel, the follow-up essential question is about a way of representing imputed results.

In this study, to address this timely and burning question, we propose a novel direction of study in data imputation where the results of imputation are formalized in the language of

* Corresponding author at: School of Electro-Mechanical Engineering, Xidian University, Xi'an 710071, PR China.
E-mail addresses: cfuzhong@gmail.com (C. Zhong), wpedrycz@ualberta.ca (W. Pedrycz), zhwli@xidian.edu.cn (Z. Li).

http://dx.doi.org/10.1016/j.asoc.2016.05.006
1568-4946/

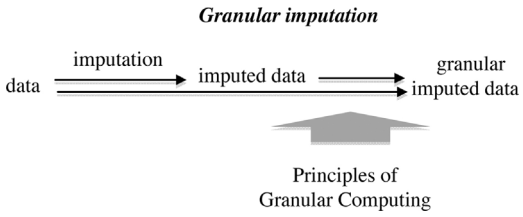

Fig. 1. A two-stage granular imputation process.

information granules [14] while the pertinent processing is built upon the principles of Granular Computing [30]. Imputed results come in the form of information granules (say, intervals) and a level of information granularity serves as a tangible indicator of the quality of the imputation process.

The main objective of this study is to introduce and motivate the usage of information granularity as a vehicle to effectively capture imputed data in the form of information granules and bring a notion of information granularity as an important asset to describe (characterize) imputed data. To the best of our knowledge, this study offers an original direction of investigations in data imputation, which has not been discussed in the past. The proposed algorithms of designing information granules using the fundamentals of Granular Computing are also original. The role of information granularity and ensuing information granules in the context of the study of imputation of data is two-fold. First, we augment the concept of imputed result by stressing that it is of different quality than the originally existing numeric data. This inherently different nature of the imputed result vis-à-vis the original numeric data becomes flagged by its granular character. Second, the produced granular result effectively quantifies the quality of the imputation mechanism being used. This is done by looking at the information granule serving as the imputed data and quantifying its level of granularity. In general, the larger (less specific) the obtained information granule, the lower the quality of the imputed result. In other words, the granular nature of the imputed results delivers a flagging effect of the quality of the imputation process and its effectiveness in the presence of available data.

To visualize a nature of the process of granular imputation, we refer to Fig. 1. It becomes apparent that the proposed approach realizes a certain follow-up process by building up on any existing numerically inclined imputation technique.

It is worth emphasizing that the proposed approach realizes a two-stage process as displayed in Fig. 1. In this sense, it can be sought as an essential augmentation (enrichment) of any imputation mechanism available in the literature by invoking the principles of Granular Computing. Furthermore we directly exploit the usage of information granules. Two main methods are investigated. In the first one, the imputation mechanism is realized for the individual variables by engaging the principle of justifiable granularity. The second one invokes the methods of fuzzy clustering, especially Fuzzy C-Means (FCM) by making the imputation method relying on all variables when imputing missing entries.

In the sequel, the granular results of imputation can be used in the construction of ensuing constructs such as classifiers, predictors and alike however by making provisions for coping of granular data (giving rise to granular classifiers, granular predictors, etc.).

The study is structured as follows. We start with some necessary prerequisites to make the material self-contained and provide all required material on Granular Computing and the principle of justifiable granularity, in particular (Section 2). In Section 3, a two-phase development of granular results of imputation is discussed. A characterization of the quality of granular imputation is discussed in Section 4 where we present the notions of coverage and specificity of produced information granules along with a synthetic view being expressed through an area of the curve (AUC) and shown in the coverage-specificity coordinates. Granular imputation realized with the aid of fuzzy clustering and an allocation of information granularity is outlined in Section 5. Experimental studies are reported in Section 6. Design aspects of granular models arising in the realm of imputed data are discussed in Section 7.

Throughout the study, we adhere to a standard notation. The data points in the collection of N data defined in the n-dimensional space of real numbers are denoted by vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k, \ldots, \boldsymbol{x}_N$, where $\boldsymbol{x}_k \in \mathbf{R}^n$, $k = 1, 2, \ldots, N$. As the data are incomplete, we introduce a Boolean matrix $\boldsymbol{B} = [b_{kj}]$, $k = 1, 2, \ldots, N$; $j = 1, 2, \ldots, n$, to capture information about the missing data. In this matrix the kj-th entry is set to zero, i.e., $b_{kj} = 0$ if the j-th variable of the k-th data point is missing otherwise for the available data the corresponding entry of the matrix is set to 1.

2. Information granules and their design—the principle of justifiable granularity

Information granules are collections of entities brought together because of some similarity, resemblance or closeness of such entities. Information granules arise as a vehicle to facilitate a description of phenomena and organizing knowledge about external world. Information granules support abstraction mechanisms. Granular Computing is about forming information granules, supporting their processing, and delivering sound interpretation vehicles [14]. There are different formal frameworks in which information granules can be described and processed, say intervals, fuzzy sets, rough sets, shadowed sets, probabilities, random sets, etc., each of them coming with their formal apparatus. Granular Computing builds a coherent setting whose principles are made general enough to apply equally well to the specific formalizations of information granules. The principle of justifiable granularity [19] delivers a conceptual and algorithmic vehicle to design an information granule on a basis of some experimental evidence (experimental data). In essence, the principle states that any information granule is formed on a basis of available existing experimental evidence where we strive that this information granule "covers" (represents) as many pieces of evidence as possible (so it is legitimized in this way) while at the same time we make it as specific (detailed) as possible while making it semantically meaningful.

In what follows, we discuss a certain specific version of the weighted version of the principle, which is of immediate relevance to this study. Let us consider that some experimental data come in the form of pairs $\boldsymbol{Z} = \{(z_1, w_1), (z_2, w_2) \ldots (z_N, w_N)\}$ where the weights $w_1, w_2, \ldots, w_N$ assuming values in the unit interval express levels of relevance (credibility) of the corresponding numeric data $z_1, z_2, \ldots, z_N$ in the construction of a certain information granule. We start by forming a numeric representative of these data. Here a weighted median comes as a sound alternative given the robustness character of the median. The weighted median med is obtained by minimizing the following expression

$$Q(\text{med}) = \sum_{k=1}^{M} w_k |x_k - \text{med}| \tag{1}$$

The minimization is realized by determining the value of the above sum by sweeping through the data points, namely $\text{med} = z_1$, $\text{med} = z_2$, …, $\text{med} = z_M$ and choosing the one which leads to the minimum of Q.

Now let us consider the ordered data (x_k, w_k), $k = 1, \ldots, N$, being a subset of the above weighted data coming in the form $\text{med} < x_1 < x_2 < \ldots < x_N$. Here we assume that the weights w_k are positive (if some data point comes with a zero weight, it is not included