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.

© 2016 Elsevier B.V. All rights reserved.

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/© 2016 Elsevier B.V. All rights reserved.

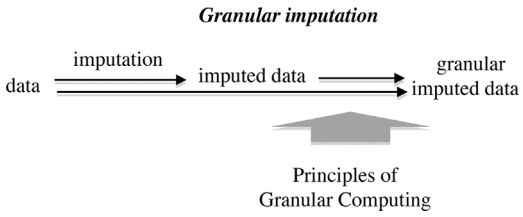

Fig. 1. A two-stage granular imputation process.

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

here. If there are some identical data, say $z_i = z_j$, we record it once with the cumulative weight equal to $w_j + w_i$).

The principle of justifiable granularity completed for interval-like information granule $[a, b]$ where $a < \text{med} < b$ is realized by forming a sound tradeoff between the requirements of experimental evidence and the specificity. Let us look at the optimization of the upper bound of the interval (b)—a construction of the lower bound is realized in an analogous manner.

The optimization criterion used in the principle is expressed as follows

$$V(b) = \left(\sum_{k:\text{med}<x_k\le b} w_k\right)\exp(-\alpha|\text{med} - b|) \tag{2}$$

The first term of the above expression quantifies the experimental evidence supporting the interval. The second term expresses the requirement of specificity. Note that by increasing the values of b, the first term increases however the second one becomes smaller. In virtue of the form assumed by V, the optimal value of b can be determined by choosing the value of b as one of the data points $x_1, x_2, \ldots, x_N$, namely $V(x_1), V(x_2), \ldots, V(x_N)$ for which V attains its maximal value. The non-negative parameter α augments the principle with some flexibility by facilitating a way of articulating the importance of the specificity requirement. For $\alpha = 0$ the specificity requirement becomes irrelevant. The increasing values of α emphasize the importance of the specificity. With this regard, it becomes beneficial to determine a range of feasible values that this parameter can assume. Obviously, the range depends upon the values of the data. Let us proceed with the determination of the upper bound of α, $\alpha_{\max}$. Here we require determining such a value of α such that the following system of inequalities is satisfied

$$\begin{aligned}
&w_1\exp(-\alpha|\text{med} - x_1|) > (w_1 + w_2)\exp(-\alpha|\text{med} - x_2|)\\
&w_1\exp(-\alpha|\text{med} - x_1|) > (w_1 + w_2 + w_3)\exp(-\alpha|\text{med} - x_3|)\\
&\ldots\\
&w_1\exp(-\alpha|\text{med} - x_1|) > (w_1 + w_2 + \ldots + w_N)\exp(-\alpha|\text{med} - x_N|)
\end{aligned} \tag{3}$$

The lower bound of α, $\alpha_{\min}$, is as a solution to the following set of inequalities

$$\begin{aligned}
&(w_1+w_2+\ldots+w_N)\exp(-\alpha|\text{med} - x_N|) > (w_1+w_2+\ldots+w_{N-1})\exp(-\alpha|\text{med} - x_{N-1}|)\\
&(w_1 + w_2 + \ldots+w_N)\exp(-\alpha|\text{med} - x_N|) > (w_1+w_2+\ldots+w_{N-2})\exp(-\alpha|\text{med} - x_{N-2}|)\\
&\ldots\\
&(w_1 + w_2 + \ldots + w_N)\exp(-\alpha|\text{med} - x_N|) > (w_1)\exp(-\alpha|\text{med} - x_1|)
\end{aligned} \tag{4}$$

As a result, we produce the range $[\alpha_{\min}, \alpha_{\max}]$, which through a linear transformation is converted into the unit interval, that is $(\alpha - \alpha_{\min})/(\alpha_{\max} - \alpha_{\min})$.

For the determination of the optimal lower bound of the interval (a), we realize a similar construct. Here we start with a set of data (y_k, w_k), $k = 1, \ldots, P$, being a subset of the above weighted data which satisfies the following property $y_P < \ldots < y_2 < y_1 < \text{med}$. As before we assume that the weights are positive and for identical data we determine the cumulative weight. By adhering to the same line of thought as discussed above we end up with the optimal lower bound (a) and the range of admissible values of α.

All in all, as a result we can generate a series of intervals $[a, b]$ indexed by a positioned in $[0, 1]$. Note that for $\alpha = 1$, we have an interval $[y_1, x_1]$ (which can be sought as a core of the fuzzy set built on a basis of a series of α-cuts).

As noted above, it is worth stressing that the principle is not confined to the formation of information granules in the form of intervals but applies equally well to other information granules such as fuzzy sets (as a matter of fact, by varying the values of α, we build the corresponding α-cuts of a certain fuzzy set so that the fuzzy set is effectively formed through the use of the representation theorem [9]). Furthermore the experimental evidence needs not to be numeric and we can envision here information granules themselves.

3. Granular data imputation—a two-phase development process

The two-phase process of data imputation consists of two main phases, namely, (i) invoking a method of imputing numeric data, and (ii) building information granules for the numeric imputations realized during the first phase. We elaborate on the details by considering a simple imputation method dealing separately with each column (variable). In what follows we determine two statistics for the individual variables computed on a basis of the available data, namely

$$m_l = \frac{\sum_{k=1}^{N} x_l b_{kl}}{\sum_{k=1}^{N} b_{kl}} \tag{5}$$

$$\sigma_l^2 = \frac{\sum_{k=1}^{N} (x_l - m_l)^2 b_{kl}}{\sum_{k=1}^{N} b_{kl}} \tag{6}$$

where $l = 1, 2, \ldots, n$. Then the average (Eq. (5)) is sought as an imputed value for the missing values in the l-th column. Let us proceed with the second phase of granular imputation, by building information granules around the numeric imputed values. It is worth stressing that the method used in the first phase takes into account only a single variable and all possible relationships among other variables are not considered at all. When moving with the second phase, we introduce some improvement of the generic imputation method by looking at possible dependencies between the data $\boldsymbol{x}$ for which we do imputation and any other data vector, say $\boldsymbol{z}$. We determine the distance between these two vectors as follows

$$\rho(\boldsymbol{x}, \boldsymbol{z}) = \sum_{l=1}^{n} \frac{(x_l - z_l)^2}{\sigma_l^2} \tag{7}$$

where σ_l is a standard deviation of the l-th variable. Obviously, the calculations above are realized for the coordinates of the vectors for which the values of both entries (x_l and z_l) are available.

We normalize the values of these distances across all data for a given $\boldsymbol{x}$ thus arriving at the following expression

$$\rho'(\boldsymbol{x}, \boldsymbol{z}_k) = \frac{\rho(x, z_k) - \min_l \rho(x, z_l)}{\max_l \rho(x, z_l) - \min_l \rho(x, z_l)} \tag{8}$$

where $x \neq \boldsymbol{z_k}$, $k = 1, 2, \ldots, N$. Subsequently we construct the weight associated with $\boldsymbol{x}$ expressing an extent to which $\boldsymbol{x}$ can be associated with $\boldsymbol{z}$ or in other words we can use $\boldsymbol{z}$ to determine eventual missing coordinates of $\boldsymbol{x}$.

$$w(\text{x}, \text{z}) = 1 - \rho'(\text{x}, \text{z}) \tag{9}$$

The higher the value of this weight, the more visible the association between $\boldsymbol{x}$ and $\boldsymbol{z}$ becomes.

Having all these prerequisites in place, we outline an overall procedure of data imputation leading to a granular result of this imputation. We consider the j-th variable for which a numeric

imputed value is m_j (this could be a median, mean or any sound numeric representative). Around this numeric m_j we form a granular imputed value by invoking the principle of justifiable granularity as presented above. The data used for the formation of the information granule come in the form of the pairs (data, weight) with the weights determined using (Eq. (9)), namely

$$x_{1j}, w(x, x_1), x_{2j}, w(x, x_2), \ldots, x_{Nj}, w(x, x_N) \tag{10}$$

where $\boldsymbol{x}$ is a data vector for which the j-th variable is subject to imputation. The bounds of the imputed interval are obtained by maximizing Q; the process is realized for the lower and upper bound, respectively.

4. Characterization of quality of granular imputation

The results of imputation are information granules and in this case the quality of imputation can be assessed from two points of view (a) the relevance of the imputed information granules, and (b) quality of the information granules. The first aspect is quantified by counting how many times information granules "cover" the missing values. Considering that p of $N \times n$ entries of data are missing, the ratio expressed in the form

$$\text{coverage} = \frac{r}{pNn} \tag{11}$$

where r is the number of cases where the imputed interval covers the missing value. The quality of the imputed intervals is quantified by defining specificity of these intervals,

$$\text{specificity} = 1 - \frac{1}{pNn}\sum_{i=1}^{pNn} \text{norm\_length}_i \tag{12}$$

where the normalized length, norm_length_i, is a ratio of the length of the imputed interval normalized by the range of the variable for which the imputation has been realized. The shorter the intervals, the higher the specificity of the resulting information granule.

While the above indicators are of a global nature (where there is no distinction among individual variables), we can look at the assessment of imputation realized for the individual variable. Considering the j-th variable where there are "t" missing values, the above measures are modified to read as follows coverage = r/t and

$$\text{specificity}_j = 1 - \frac{1}{t\sum_{i=1}^{t} \text{norm\_length}_i} \tag{13}$$

Notably, both the coverage and specificity are functions of α. The first one is a non-increasing function of α whereas the second descriptor, specificity, is a non-decreasing function of α. A compromise between these two characteristics can be achieved by selecting a suitable value of α. A global scalar quantification of the quality is formed by computing the area under curve (AUC) obtained in the coverage-specificity coordinates and indexed by successive values of α.

5. Imputation with the use of fuzzy clustering and its granular augmentation

Fuzzy clustering, say Fuzzy C-Means [3,15] is a generic vehicle to reveal a structure in data. In light of this observation, once the clustering has been completed for all available complete data, we can impute the missing values. Fuzzy clustering is regarded as a fundamental way of forming information granules on a basis of experimental data and in this way reveals the underlying structure present in the data. In the context of data imputation, there are some benefits: (i) all variables and relationships among them are taken into account when building representatives (prototypes) of the data so that (or and hence) these prototypes can be effectively used in the imputation method, (ii) clustering builds some general dependencies and we are not confined to the detailed functional relationships (that might be difficult to verify), which are behind some sophisticated imputation techniques. Some studies on fuzzy clustering and imputation are reported in Refs. [2,5,8,10,27].

In what follows, we briefly review the method as it is considered in the setting of incomplete data. The objective function shown below and guiding a process of formation of clusters is expressed as a sum of distances between data and prototypes where $v_1, v_2, \ldots, v_c$ are the prototypes of the clusters and $\boldsymbol{U} = [u_{ik}]$ is a partition matrix while m, $m > 1$, is a fuzzification coefficient impacting a geometry of the membership functions (entries of the partition matrix).

$$Q = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{ik}^m ||x_k - v_i||_B^2 \tag{14}$$

The distance $||.||$ standing in Eq. (14) is a weighted Euclidean distance function expressed as

$$||x_k - v_i||_B^2 = \sum_{j}^{n} \frac{(x_{kj} - v_{ij})^2}{\sigma_j^2} b_{kj} \tag{15}$$

where σ_j is a standard deviation of the j-th variable. Note that the Boolean matrix $\boldsymbol{B}$ used in the above calculations emphasizes that only available data are involved in the computations.

The FCM method returns a collection of prototypes and the partition matrix, which are determined in an iterative fashion using the following formulas

$$\begin{aligned} u_{ik} &= \frac{1}{\sum_{j=1}^{c} \left((||x_k - v_i||_B) / \left(||x_k - v_j||_B \right) \right)^{2/(m-1)}} \\ v_{ij} &= \frac{\sum_{k=1}^{N} u_{ik}^m x_{kj} b_{kj}}{\sum_{k=1}^{N} u_{ik}^m b_{kj}} \end{aligned} \tag{16}$$

where $i = 1, 2, \ldots, c$; $j = 1, 2, \ldots, n$; $k = 1, 2, \ldots, N$.

Note that these are modified expressions used originally in the FCM; here in the computations we eliminate the missing entries (using the Boolean values of $\boldsymbol{B}$).

Now a certain input $\boldsymbol{x}$ whose some values are missing becomes reconstructed viz. the missing values are imputed. Let us associate with $\boldsymbol{x}$ a Boolean vector $\boldsymbol{b}$ whose values set to 0 indicate that the corresponding entry (variable) is missing. The missing values are imputed as a result of a two-phase process:

(i) determination of membership grades computed on a basis of the prototypes and using only available inputs of $\boldsymbol{x}$

$$\tilde{u}_i = \frac{1}{\sum_{j=1}^{c} \left((||x_k - v_i||_B) / \left(||x_k - v_j||_B \right) \right)^{2/(m-1)}} \tag{17}$$

(ii) computing the missing entries of $\boldsymbol{x}$ coming as a result of a reconstruction process [16]

$$\tilde{x}_j = \frac{\sum_{i=1}^{c} \tilde{u}_i^m v_{ij}}{\sum_{i=1}^{c} \tilde{u}_i^m} \quad (18)$$

where the indexes "j" are those for which the entries of $\boldsymbol{b}$ are equal to 0.

The quality of the imputation process realized with the use of fuzzy clustering is expressed by means of the following sum

$$F = \frac{1}{\text{number of missing data}} \sum_{\substack{k,i: \\ b_{ik}=0}} \frac{(x_{kj} - \hat{x}_{kj})^2}{\sigma_j^2} \quad (19)$$

that can be referred to as an imputation error. Note that this error is determined with respect to all missing data. Along with Eq. (19) we can compute the following expression

$$G = \frac{1}{\text{number of available data}} \sum_{\substack{k,i: \\ b_{ik}=1}} \frac{(x_{kj} - \hat{x}_{kj})^2}{\sigma_j^2} \quad (20)$$

where the above sum is taken over all entries of the data that were originally available. Obviously, in this case $\tilde{x}_j$ is computed for indexes for which the entries of $\boldsymbol{b}$ are equal to 1. This is referred to as a reconstruction error that is inherently associated with the process of granulation–degranulation (or encoding–decoding) as discussed in Ref. [16].

To make the numeric results of imputation granular, we consider a mechanism of allocation of information granularity, viz. a way of forming intervals around the imputation results [17,18]. Just to note that we do not apply the principle of justifiable granularity as the number of clusters (prototypes) is quite low, therefore forming an information granule in the presence of a few data points might not be well justified. The intervals constructed symmetrically around each numeric imputed result come in the form

$$\hat{X}_{kj} = [\hat{x}_{kj} - \frac{\varepsilon}{2}\text{range}_j, \hat{x}_{kj} + \frac{\varepsilon}{2}\text{range}_j] \quad (21)$$

where ε is a parameter controlling a level of allocated information granularity, $\varepsilon \in [0,1]$ while range_j is the range of values assumed by the j-th variable. The higher the value of ε, the broader the obtained interval and the lower its specificity.

The above construction of granular results can be made more flexible by admitting an asymmetric distribution of the interval around the imputed value. Here we have

$$\hat{X}_{kj} = [\hat{x}_{kj} - \varepsilon\gamma\text{range}_j, \hat{x}_{kj} + \varepsilon(1-\gamma)\text{range}_j] \quad (22)$$

where $\gamma \in [0,1]$ is an index of asymmetry. Obviously the previously formed intervals (Eq. (21)) are special cases of those formed by using Eq. (22) when $\gamma = 1/2$.

As before the quality of granular imputation is expressed by means of the coverage and specificity criteria.

6. Experimental studies

In the following series of experiments, we quantify the performance of the proposed imputation methods and offer a thorough comparative analysis. As the results come in the form of interval-valued information granules, the evaluation of the method is carried out by inspecting coverage-specificity characteristics (whose visualization offers an insight into possible tradeoffs between these two descriptors) and reporting the global characterization of the method in the form of the AUC values. Furthermore we report results of parametric analysis by investigating an impact of the key parameters (such as p, m, and c) on the performance of the method. Along with the evaluation completed for the granular outcomes of the imputation (which is unique one and as such not comparable with the numeric results delivered by other imputation techniques known in the literature), the experiments are completed for some numeric-outcome generating imputation methods such as average-based and FCM-based imputation.

Table 1
Data sets used in experimental studies.

Data set	Number of data	Dimensionality
Housing	506	14
Climate model simulation crashes	540	18
Blood transfusion service center	748	5
Airfoil self-noise	1503	6

Table 2
AUC values produced for selected values of p.

Data set	p					
	0.05	0.10	0.20	0.30	0.40	0.45
Housing	0.4436	0.4569	0.4548	0.4604	0.4620	0.4617
Climate	0.4770	0.4834	0.4828	0.4812	0.4797	0.4811
Blood	0.3929	0.4019	0.3844	0.3887	0.3763	0.3784
Airfoil	0.3408	0.3386	0.3447	0.3370	0.3411	0.3353

We consider a collection of data coming from the Machine Learning repository http://archive.ics.uci.edu/ml/datasets.html, see Table 1.

We remove p of all inputs of the data that is pNn randomly and for such incomplete data set we carry out the imputation procedure. The corresponding values of p are set as 5, 10, 15, . . ., 45%.

6.1. Granular imputation based on mean of individual variables

We show the results in terms of the coverage treated as a function of α and the specificity–coverage characteristics; refer to Fig. 2.

Not surprising, the coverage is a decreasing function of α. The character of this relationship is interesting and it varies from data to data. In several cases we observe some plateau regions, as e.g., in Housing, Blood, and Airfoil whereas for some other, namely climate there is a rapid drop in the coverage values. These regions point at some suitable values of α one can select. For instance, in the Housing data, the coverage drops at $\alpha = 0.1$ and stabilizes further meaning that one can make the intervals quite specific by pushing the value of α up to 0.9 and making the intervals more specific. These characteristics offer an interesting insight into the abilities of the granular imputed results to represent the data. Noticeably the coverage does not depend visibly upon the values of p as the curves overlap.

The overall characterization of the quality of granular imputation expressed in terms of the AUC values is provided in Table 2. Again the characteristics look different for different data sets and the shapes of the curves are different. Furthermore the curves help identify a "knee" points pointing at preferred value of α. This happens for the Housing ($\alpha = 0.7$), Blood and Airfoil. In some cases several knee points are observed. The AUC values could offer a general view at the performance of this imputation method on the data set; here the best performance is reported for the two first data sets (0.4436 and 0.4770) while lower performance is reported for the two other data sets with the weakest result obtained for the Airfoil.

6.2. Granular imputation based on fuzzy clustering

All experiments were carried out in the same manner as before. The number of clusters (c) and the fuzzification coefficient (m) are

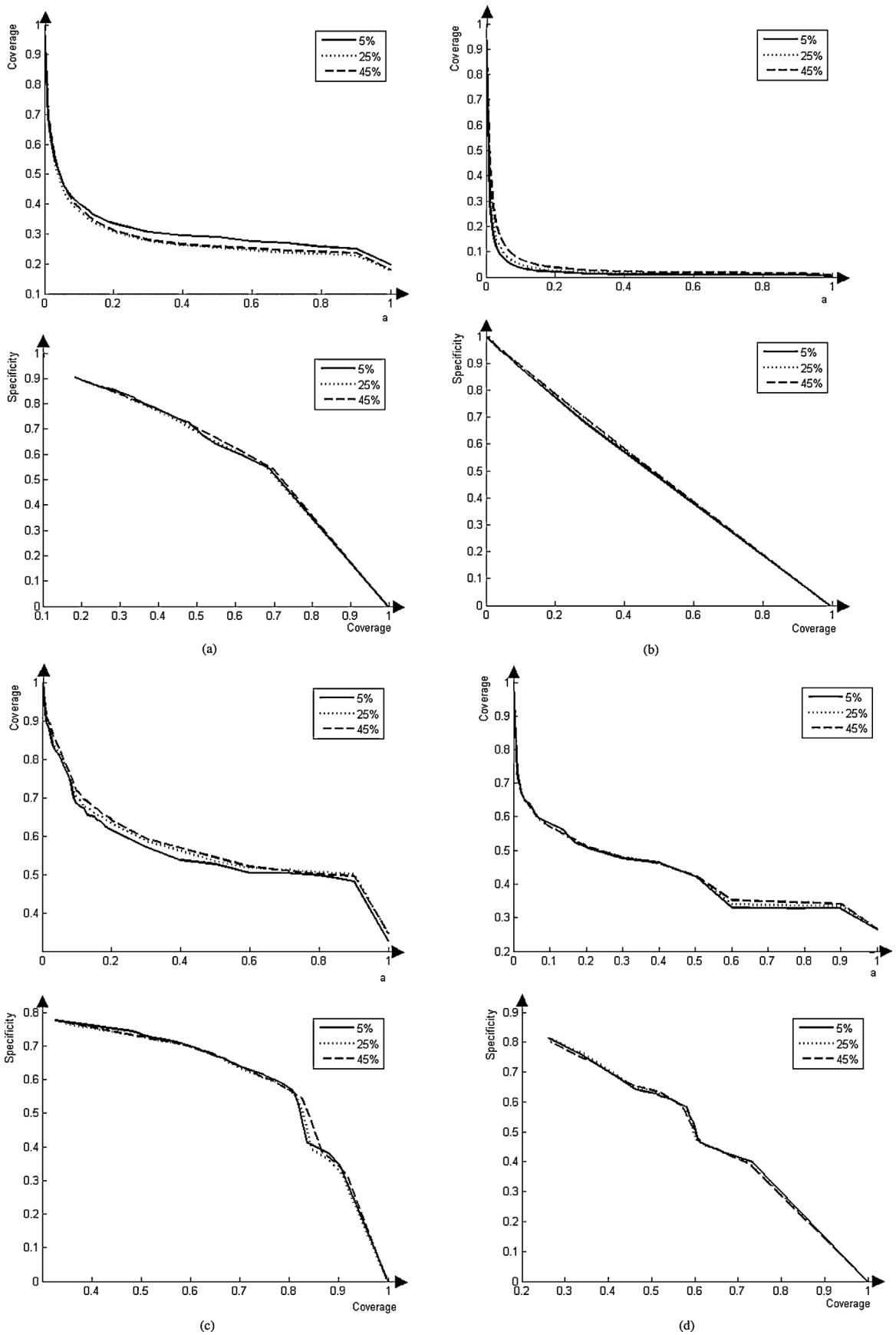

Fig. 2. Specificity–coverage characteristics for Machine Learning data sets for $p = 5\%$, 25, and 45%; (a) Housing, (b) Climate, (c) Blood, and (d) Airfoil.

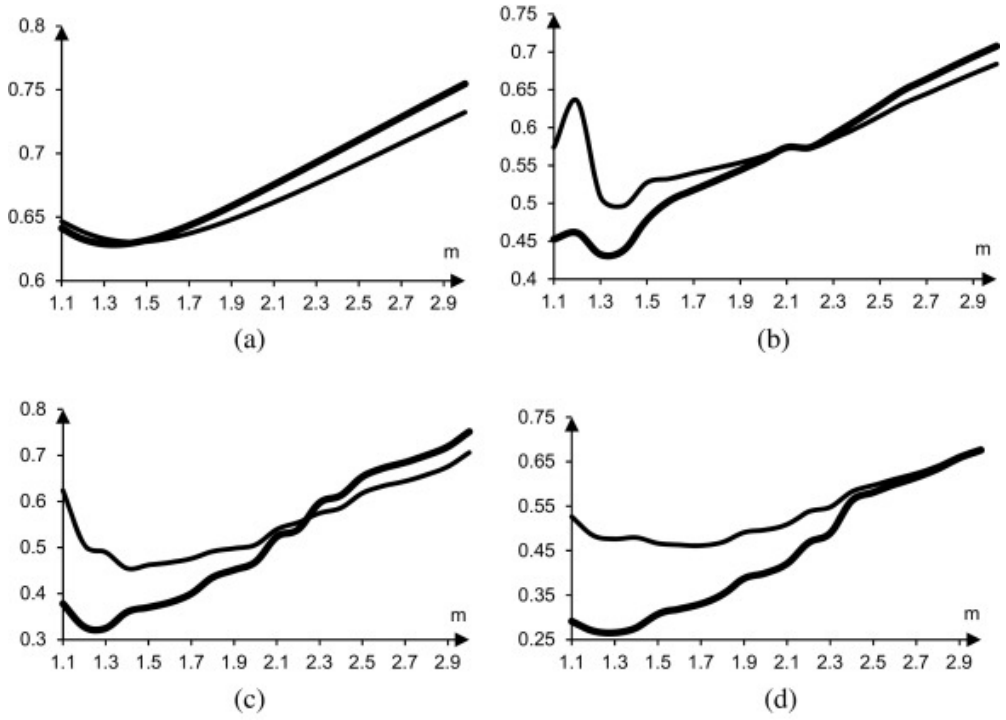

Fig. 3. Plots of reconstruction error (G) and imputation error (F) regarded as a function of m for selected number of clusters: $c = 2$ (a), $c = 5$ (b), $c = 9$ (c), and $c = 13$ (d). The level of missing data (p) is 0.20; thick line—reconstruction error, line—imputation error.

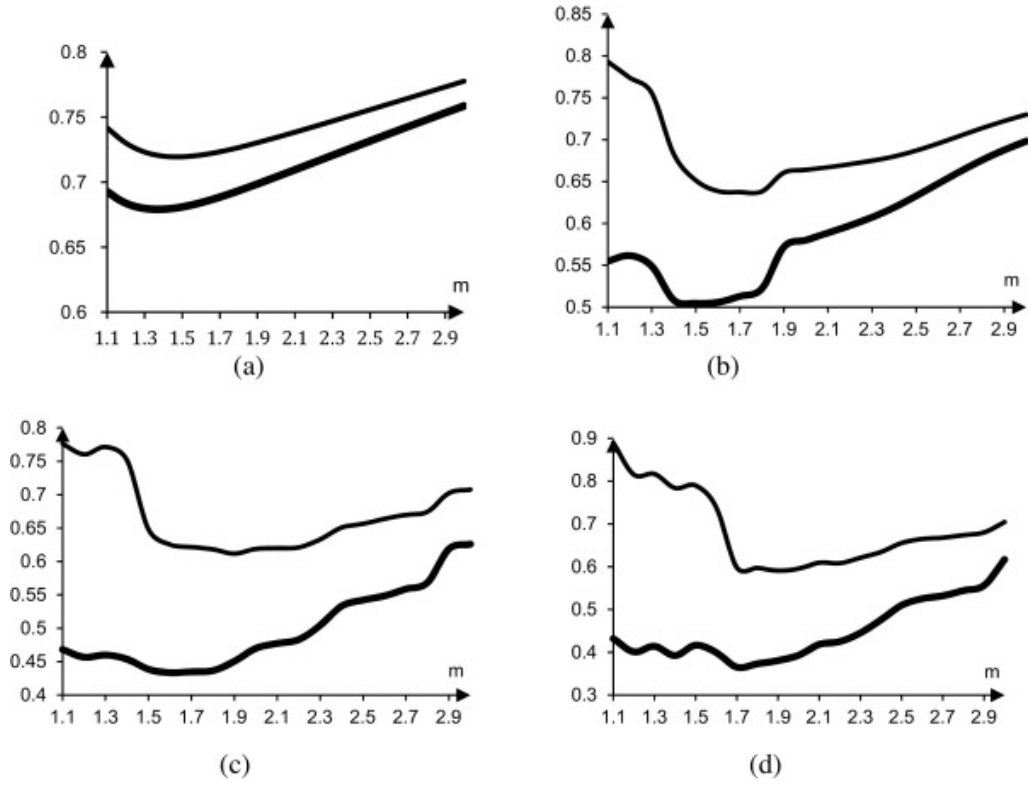

Fig. 4. Plots of reconstruction error (G) and imputation error (F) regarded as a function of "m" for selected number of clusters: $c = 2$ (a), $c = 5$ (b), $c = 9$ (c), and $c = 13$ (d). The level of missing data (p) is 0.45; thick line—reconstruction error, line—imputation error.

the two essential adjustable parameters whose values impact the performance of the imputation results. We also consider several levels of missing data, say $p = 0.05$, 0.10, 0.20, 0.30, and 0.40.

Housing data Considering a certain level of missing data, fuzzy clustering has been carried out for selected number of information granules, namely $c = 2$, 5, 9, and 13. These numbers were selected in a way so that the performance of the imputation mechanism can be

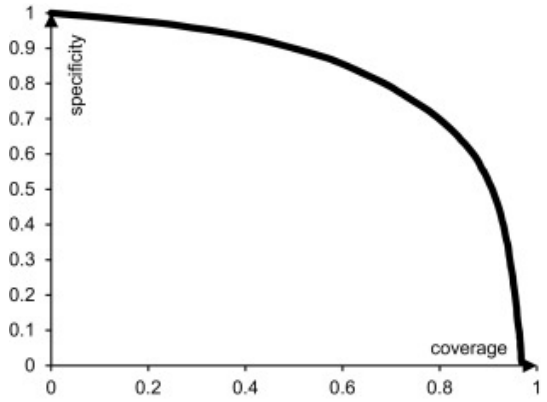

Fig. 5. Coverage–specificity characteristics of the imputation process; the results shown for the optimal values of m, c and $p = 0.40$.

Table 3
Imputation error F for the granular imputation (FCM-based) and based on the averages—Housing.

p	FCM-based imputation	Average-based imputation
0.05	0.441	0.965
0.10	0.394	0.959
0.20	0.453	0.979
0.30	0.577	1.082
0.40	0.542	1.017
0.45	0.591	1.048

observed for a broad range of the number of information granules. For comparative reasons, we also look at the reconstruction error and compare it with the produced imputation error. In the series of plots, Figs. 3 and 4, these two errors are reported as function of the fuzzification coefficient. The two quite different values of p help contrast the behavior of the method in terms of the reconstruction and imputation capabilities.

The plots of the characterization of the granular realizations of the imputed values provide an insight into the dependency between the coverage of the data and the specificity of the produced results. For the optimal values of m and c, the corresponding coverage–specificity dependencies are shown in Fig. 5.

For the optimal values of m and c, we display the AUC for symmetric and asymmetric allocation of information granularity for several values of p (Fig. 6).

The AUC values are lower when p increases. For instance, if $c = 13$, the AUC for $p = 0.20$ is 0.82 while for $p = 0.45$ we have AUC of 0.79. For comparison, we also report the imputation error in case the missing data are replaced by the averages standing in the corresponding columns of the data set, see Table 3.

The results obtained for the remaining data sets are summarized in terms of the imputation error obtained for the optimal values of m and c, the corresponding values of the AUC index for the symmetric and asymmetric intervals, see Table 4. For reference, provided are the values of the imputation error in case the imputation realized in terms of the mean for each variable.

Based on the results shown in this table, the imputation error exhibits an increasing tendency for higher values of p. The decreasing tendency is observed for the AUC values. The improvement offered by the asymmetric intervals over the results formed for the symmetric ones is present but the differences are quite limited. The improvement of the FCM-based imputation over the plain mean-based imputation done for the individual variables is highly visible, for several data sets (Housing and Blood) reducing the imputation error more around two times.

7. Granular imputation and a construction of fuzzy models

Imputed data are granular and as such they have to be treated when constructed fuzzy models or any models. Depending on the location of the imputed data, we discuss four alternatives, which are to be treated in the development (estimation of parameters) of the ensuing models. The performance index has to be modified to cope with granular data. In what follows, we use capital letters to denote the imputed data. Denote the model by M, so for any input $\boldsymbol{x}$, the model returns $y = M(\boldsymbol{x})$.

Distinguished are four alternatives depending on the nature of the imputed data:

1 $(\boldsymbol{x}_k, \text{target}_k)$—no imputation; there are original numeric data and subsequently there is no change to the typical performance index used in the construction of numeric models. It could be a commonly encountered RMSE index, say

$$\text{RMSE} = \sqrt{\frac{1}{M_1}\sum_{k=1}^{M_1}(M(\boldsymbol{x}_k) - \text{target}_k)^2} \tag{23}$$

where M_1 is the number of data for which no imputation was applied.

2 $(\boldsymbol{X}_k, \text{target}_k)$ imputation of the input data (concerning one or several input variables). The granular input $\boldsymbol{X}_k$ implies that the result of the model is granular as well, $Y_k = M(\boldsymbol{X}_k)$. In this regard, one evaluates an extent to which target is included (contained) in Y_k, say $\text{incl}(\text{target}_k, Y_k)$. The measure of inclusion can be specialized depending upon the nature of information granule of the imputed data. In case of sets (intervals), the concept is binary and relates to the inclusion predicate

$$\text{incl}(\text{target}_k, Y_k) = \begin{cases} 1 & \text{if } \text{target}_k \in Y_k \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

In case of fuzzy sets (Y_k being described by some membership function), the inclusion predicate returns a degree of membership of target_k in Y_k, viz. $Y_k(\text{target}_k)$.

3 $(\boldsymbol{x}_k, \text{Target}_k)$ imputation realized for the output data. The output of the model is numeric, $y_k = M(\boldsymbol{x}_k)$ and we compare it with the granular target, Target_k. As before we use the same way of evaluation of the quality of the model as discussed in the previous scenario.

4 (X_k, Target_k) imputation involves both the input and output variables. The output of the model is granular Y_k, which has to be compared with Target_k. Here a measure of comparing (matching) two information granules is required.

With regard to this situation, let us start with Y_k and Target_k being two intervals, namely $[y_k^-, y_k^+]$ and $[\text{target}_k^-, \text{target}_k^+]$. For them we introduce two operations, join and meet, expressed as follows

$$\begin{aligned} \text{join} \quad & Y_k \oplus \text{Target}_k = \left[\min(y_k^-, \text{target}_k^-), \max(y_k^+, \text{target}_k^+)\right] \\ \text{meet} \quad & Y_k \oplus \text{Target}_k = \left[\max(y_k^-, \text{target}_k^-), \min(y_k^+, \text{target}_k^+)\right] \end{aligned} \tag{25}$$

For the meet operation, we consider that Y_k and Target_k are not disjoint; otherwise the meet is empty. The meaning of these two operations is clarified in Fig. 7.

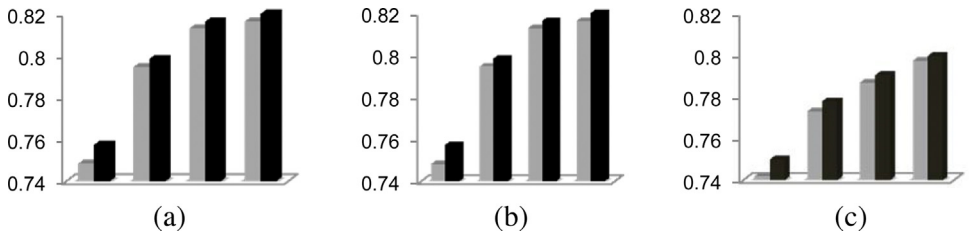

Fig. 6. AUC values—symmetric (grey color bar) and asymmetric allocation of information granularity for $c = 2, 5, 9$, and 13 (bars from left to right) for selected values of p: (a) $p = 0.20$, (b) $p = 0.30$, and (c) $p = 0.45$.

Table 4
Results of imputation process obtained for several Machine Leaning data sets; shown are the optimal parameters of the clustering method, obtained imputation error, the AUC values, and the mean-based imputation.

Climate						
p	c_{opt}	m_{opt}	Imputation error	AUC-symmetr	AUC-asymmetr	Mean-based imputation error
0.05	10	1.2	1.2885	0.893	0.903	1.472
0.10	11	1.2	1.0689	0.905	0.914	1.148
0.20	4	1.2	1.2762	0.899	0.909	1.312
0.30	2	1.5	1.3249	0.904	0.913	1.337
0.40	2	1.6	1.3246	0.904	0.904	1.338
0.45	2	1.6	1.1934	0.904	0.913	1.206
Blood						
p	c_{opt}	m_{opt}	Imputation error	AUC-symmetr	AUC-asymmetr	Mean-based imputation error
0.05	11	2.1	0.3970	0.851	0.864	0.728
0.10	13	1.8	0.4672	0.845	0.851	0.925
0.20	13	2.0	0.4875	0.831	0.837	0.892
0.30	13	2.0	0.5560	0.824	0.829	0.904
Airfoil						
p	c_{opt}	m_{opt}	Imputation error	AUC-symmetr	AUC-asymmetr	Mean-based imputation error
0.05	12	1.9	0.5591	0.699	0.709	0.873
0.10	11	2.0	0.6092	0.686	0.694	0.961
0.20	11	2.0	0.6221	0.671	0.679	0.960
0.30	13	2.1	0.7406	0.652	0.658	0.997

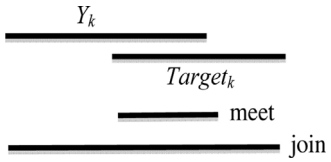

Fig. 7. Join and meet of interval information granules.

To quantify an extent to which these intervals Y_k and Target_k match, we propose the following definition of a degree of matching, $\xi(Y_k, \text{Target}_k)$, which assumes a form of the following ratio

$$\xi(A, B) = \frac{|Y_k \otimes \text{Target}_k|}{|Y_k \oplus \text{Target}_k|} \tag{26}$$

The crux of this definition is to regard a length of the meet, $|Y_k \oplus \text{Target}_k|$ as a measure of overlap of the intervals and calibrate this measure by taking the length of the join, $|Y_k \oplus \text{Target}_k|$. The essence of the above expression associates with the Jaccard's coefficient being used to quantify similarity.

If the above information granules are fuzzy sets Y_k and Target_k, the above construct can be easily used by considering their α-cuts of fuzzy sets, determining the value of the degree of matching $\xi(\ldots)$ and aggregating the results over all values of α, namely $\int_0^1 \xi(Y_{k\alpha}, \text{Target}_{k\alpha}) d\alpha$ where $Y_{k\alpha}$ and $\text{Target}_{k\alpha}$ are the corresponding α-cuts of the fuzzy sets.

Overall, the performance index comprises two components: the one, which is computed on a basis of numeric results produced by the model (scenario 1) and another one which involves inclusion or matching two information granules (scenarios 2–4). More formally, we can structure the global performance index as follows

$$Q = \sqrt{\frac{1}{M_1}\sum_{k=1}^{M_1}(M(x_k) - \text{target}_k)^2} + \beta\{\frac{1}{M_2}\sum_{k=1}^{M_2}(1 - \text{incl}(\text{target}_k, Y_k)) + \frac{1}{M_3}\sum_{k=1}^{M_3}(1 - \xi(\text{Target}_k, Y_k))\} \tag{27}$$

where β is a certain weight factor, which helps striking a sound balance between the two components of the performance index involving original data and those being subject to imputation. M_2 denotes the number of data concerning the second and the third scenario while M_3 is the number of cases dealing with the fourth scenario. The parameters of the model are optimized such that the performance index Q becomes minimized.

8. Conclusions

In this study, we discussed the problem of data imputation formulated in the new framework of Granular Computing. We showed that this two-phase approach enhances the existing techniques of imputation by making the results granular—this evidently helps tell apart the original numeric data from those being the result of imputation. Furthermore the approach becomes essential to quantify the quality of the imputed data by stressing their granular nature. The measure of specificity is crucial with this regard while the

coverage index characterizes the quality of the imputation process. The plots of the coverage–specificity relationships provide a general view at the nature and the quality of the process and can be sought as a certain high-level synthetic signature of the imputation process and the nature of the data. It is worth stressing that the developed concepts of granular imputation can be viewed as a follow-up process following any imputation technique, which speaks to its general nature and visible applicability of the introduced methodology. The value of the AUC measure computed on a basis of the coverage–specificity plot serves as a high-end indicator of the quality of the originally used imputation procedure.

The two principles of Granular Computing, namely the principle of justifiable granularity and the allocation of information granularity being regarded as a design asset serve as the underlying conceptual and algorithmic environment.

The study could be sought as an initial point of investigations in this area with several possible and potentially promising pursuits. First, the detailed investigations reported here exploit intervals as information granules. While we did this on purpose to better illustrate the key concepts, it would be interesting to look at other formalisms of information granules, especially fuzzy sets and rough sets. Given the general framework of Granular Computing, this line of study could be completed naturally as the discussed principles are independent from the formal granular setting being adopted in the procedures of granular imputation.

Interestingly, the granular nature of imputed results can be encountered when dealing with incomplete linguistic data, viz. the data whose values are information granules (say, fuzzy sets) and some of the entries are missing (expressed by fuzzy sets such as unknown). Following the principle outlined in the study, the imputed results are information of higher type than the granular data one originally started with (say, fuzzy sets of type-2). The role of granular data resulting from the imputation process can be used and exploited in the ensuing granular models (such as classifiers or predictors). The results of the models are now granular and this way effectively quantify the performance of the model by making it more in rapport with reality and reflecting the fact that it has been constructed in presence of incomplete (and then imputed) data implying a varying level of information granularity of the classification and prediction results depending upon the region of the feature space and the quality of data present there. These issues will be investigated in the future studies.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 51305325 and 61374068, the Recruitment Program of Global Experts, and the Science and Technology Development Fund, MSAR, under Grant No. 078/2015/A3.

References

[1] R. Andridge, R. Little, A review of hot deck imputation for survey non-response, Int. Stat. Rev. 78 (1) (2010) 40–64.
[2] I.B. Aydilek, A. Arslan, A hybrid method for imputation of missing values using optimized Fuzzy C-Means with support vector regression and a genetic algorithm, Inf. Sci. 233 (2013) 25–35.
[3] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[4] K.V. Branden, S. Verboven, Robust data imputation, Comput. Biol. Chem. 33 (1) (2009) 7–13.
[5] A.G. Di Nuovo, Missing data analysis with Fuzzy C-Means: a study of its application in a psychological scenario, Expert Syst. Appl. 38 (6) (2011) 6793–6797.
[6] M.A. Farhangfar, L.A. Kurgan, W. Pedrycz, A novel framework for imputation of missing values in databases, IEEE Trans. Syst. Man Cybern. A: Syst. Hum. 37 (5) (2007) 692–709.
[7] P. Gómez-Carracedo, J.M. Andrade, P. López-Mahía, S. Muniategui, D. Prada, A practical comparison of single and multiple imputation methods tohandle complex missing data in air quality datasets, Chemom. Intell. Lab. Syst. 134 (2014) 23–33.
[8] H. Ichihashi, K. Honda, A. Notsu, T. Yagi, Fuzzy C-Means classifier with deterministic initialization andmissing value imputation, Proc. of the 2007 IEEE Symposium on Foundations of Computational Intelligence (FOCI 2007) (2007) 214–221.
[9] G. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall, Upper Saddle River, 1995.
[10] D. Li, H. Gu, L.Y. Zhang, A Fuzzy C-Means clustering algorithm based on nearest neighbor intervals for incomplete data, Expert Syst. Appl. 37 (10) (2010) 6942–6947.
[11] L. Li, Y. Li, Z. Li, Missing traffic data: comparison of imputation methods, IET Intell. Transp. Syst. 8 (1) (2014) 51–57.
[12] R.J.A. Little, D.B. Rubin, Statistical Analysis with Missing Data, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2002.
[13] Y. Liu, S.D. Brown, Comparison of five iterative imputation methods for multivariate classification, Chemom. Intell. Lab. Syst. 120 (2013) 106–115.
[14] W. Pedrycz, Granular Computing Analysis and Design of Intelligent Systems, CRC Press/Francis Taylor, Boca Raton, 2013.
[15] W. Pedrycz, Knowledge-Based Fuzzy Clustering, John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.
[16] W. Pedrycz, J. Valente de Oliveira, A development of fuzzy encoding and decoding through fuzzy clustering, IEEE Trans. Instrum. Meas. 57 (4) (2008) 829–837.
[17] W. Pedrycz, Allocation of information granularity in optimization and decision-making models: towards building the foundations of Granular Computing, Eur. J. Oper. Res. 232 (1) (2014) 137–145.
[18] W. Pedrycz, From logic descriptors to granular logic descriptors: a study in allocation of information granularity, J. Ambient Intell. Humaniz. Comput. 4 (4) (2013) 411–419.
[19] W. Pedrycz, W. Homenda, Building the fundamentals of granular computing: a principle of justifiable granularity, Appl. Soft Comput. 13 (2013) 4209–4218.
[20] G. Rahman, Z. Islam, Missing value imputation using decision trees and decision forestsby splitting and merging records: two novel techniques, Knowl. Based Syst. 53 (2013) 51–65.
[21] V. Ravi, M. Krishna, A new on line data imputation method based on general regression auto associative neural network, Neurocomputing 138 (2014) 106–113.
[22] Y. Ren, G. Li, J. Zhang, W. Zhou, Lazy collaborative filtering for data sets with missing values, IEEE Trans. Cybern. 43 (6) (2013) 1822–1834.
[23] J.M. Robins, A. Rotnitzky, D.O. Scharfstein, Semiparametric regression for repeated outcomes with non-ignorable non-response, J. Am. Stat. Assoc. 93 (1998) 1321–1339.
[24] D.B. Rubin, Multiple Imputation for Non Response in Surveys, Wiley, New York, 1987.
[25] J.L. Schafer, Multiple imputation: a primer, Stat. Methods Med. Res. 8 (1999) 3–15.
[26] J.L. Schafer, J.W. Graham, Missing data: our view of the state of the art, Psychol. Methods 7 (2) (2002) 147–177.
[27] H. Timm, C. Doring, R. Kruse, Different approaches to fuzzy clustering of incomplete data sets, Int. J. Approx. Reason. 35 (3) (2004) 239–249.
[28] J. Van Hulse, T.M. Khoshgoftaar, Incomplete-case nearest neighbor imputation in software measurement data, Inf. Sci. 259 (2014) 596–610.
[29] D. Williams, X. Liao, Y. Xue, L. Carin, B. Krishnapuram, On classification with incomplete data, IEEE Trans. Pattern Anal. 29 (3) (2007) 427–436.
[30] T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern. 43 (6) (2013) 1977–1989.
[31] P. Zhang, Multiple imputation: theory and method, Int. Stat. Rev. 71 (3) (2003) 581–592.
[32] S. Zhang, Z. Jin, X. Zhu, Missing data imputation by utilizing information within incomplete instances, J. Syst. Softw. 84 (3) (2011) 452–459.