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

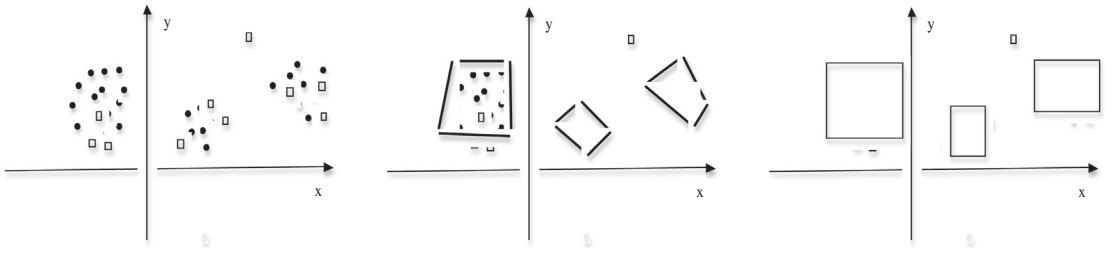

Fig. 1. Example data (black dots) to be described in a two-dimensional space; shown are also data belonging to the second class (squares) to be excluded from the descriptors formed for the first class (a) along with a collection of geometric descriptors of diverse shape (b) and of rectangular shape (c).

comes a well-developed discipline of interval analysis. Alternatively to an enumeration of elements belonging to a given set, sets are described by characteristic functions taking on values in {0, 1}.

Fuzzy sets [22] provide an important conceptual and algorithmic generalization of sets. By admitting partial membership of an element to a given information granule we bring an important feature which makes the concept to be in rapport with reality. It helps working with the notions where the principle of dichotomy is neither justified nor advantageous. The description of fuzzy sets is realized in terms of membership functions taking on values in the unit interval. Formally, a fuzzy set A is described by a membership function mapping the elements of a universe $\mathbf{X}$ to the unit interval [0, 1].

Shadowed sets [15,17] offer an interesting description of information granules by distinguishing among elements, which fully belong to the concept, are excluded from it and whose belongingness is completely *unknown*. Formally, these information granules are described as a mapping X: $\mathbf{X} \rightarrow \{1, 0, [0, 1]\}$ where the elements with the membership quantified as the entire [0, 1] interval are used to describe a shadow of the construct. Given the nature of the mapping here, shadowed sets can be sought as a granular description of fuzzy sets where the shadow is used to localize unknown membership values, which in fuzzy sets are distributed over the entire universe of discourse. Note that the shadow produces non-numeric descriptors of membership grades.

Probability-oriented information granules are expressed in the form of some probability density functions or probability functions. They capture a collection of elements resulting from some experiment. In virtue of the concept of probability, the granularity of information becomes a manifestation of occurrence of some elements. For instance, each element of a set comes with a probability density function truncated to [0, 1], which quantifies a degree of membership to the information granule.

Rough sets emphasize a roughness of description of a given concept X when being realized in terms of the indiscernibility relation provided in advance. The roughness of the description of X is manifested in terms of its lower and upper approximations of a certain rough set. One can refer to a plethora of applications [5,9].

3. Problem formulation

In what follows, we briefly formulate the problem in a formal way and then outline a general flow of processing. Given is a data set composed of M vectors located in an n-dimensional space of real numbers $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_M$ where $\boldsymbol{x}_k \ \boldsymbol{R}^n, k = 1, 2, \ldots, M$. For this data set we formulate a certain category of data $\boldsymbol{D}$ (being a subset of the entire dataset $\boldsymbol{F}$) of dimensionality $N, N < M$, which is of interest to us and whose description is of interest to us. Let elaborate on the nature of the problem discussed here. We are concerned with a two-class problem where a data set $\boldsymbol{F}$ comprises a mixture of data coming from two classes. $\boldsymbol{D}$ is a subset of $\boldsymbol{F}$, which consists of data coming only from a single class.

Our ultimate objective is to formulate a sound description of $\boldsymbol{D}$ as a collection of meaningful and interpretable (transparent) descriptors. As stressed in the previous section, to meet the requirement of relevance and interpretability, descriptors are information granules. The ensuing development process of information granules is prudently established so that the two requirements are carefully addressed. The proposed overall scheme along with its main conceptualization phases are formed with this research agenda in mind, refer to Fig. 2. The proposed flow of processing offers a certain level of originality by casting the problem in a new setting of Granular Computing and stressing a role of granular descriptors in the characterization of the interpretable structure of the data.

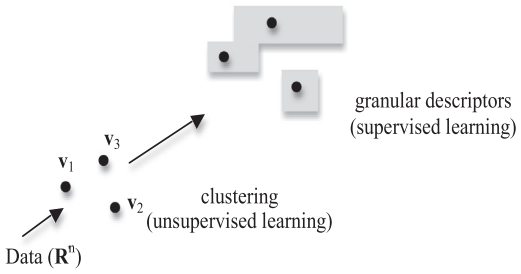

Fig. 2. From data to granular descriptors.

There are two fundamental steps used in the realization of the granular descriptors. First a structure in the data $\boldsymbol{D}$ is formed by involving various mechanisms of data clustering. Clustering and fuzzy clustering, in particular are sought as a constructive and commonly used vehicle to reveal a structure in the data. Partition-based methods governed by some objective function are a viable alternative worth exploring here. Once fuzzy clustering (say, Fuzzy C-Means, FCM, or other fuzzy sets oriented algorithms, see [4]) has been completed, the results come in the form of a collection of numeric prototypes $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c$. They can be sought as a blueprint of the descriptors. At the subsequent phase of processing, see Fig. 2, information granules are being formed around the prototypes. Here their construction is realized by the principle of justifiable granularity. By looking at the underlying processing paradigms, these two phases comes with a visible rationale. The first phase invokes unsupervised learning (clustering), is focused on the elements of $\boldsymbol{D}$ and leads to the formation of the structural blueprint of the data. The clustering strategy becomes useful as we first embark on exploiting the essence of the data and ignoring at this stage somewhat disparaging details, which are carefully handles afterwards by building a more comprehensive and well-rounded granular constructs. It is also noticeable that clustering ignores the use of data coming from $\boldsymbol{F}$–$\boldsymbol{D}$, which have to be used in the realization of information granules. The principle of justifiable granularity is completed in a supervised mode and in this way helps built information granules that capture the structural crux of $\boldsymbol{D}$.

4. Fuzzy clustering- a data-oriented view at Fuzzy C-Means

We briefly review a method of Fuzzy C-Means [2,13], which could be sought as a generic vehicle to build information granules.

In the setting of our discussion, fuzzy clustering arises here as a useful conceptual and algorithmic alternative. As a result, we arrive at a collection of information granules. We consider a collection of n-dimensional numeric data $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. Fuzzy C-Means (FCM) [2] supplies a sound mechanism of building information granules formed on a basis of a collection of numeric data. The formation of information granules is realized by minimizing an objective function expressing a spread of data around prototypes

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

where c stands for the number of clusters. The description of the clusters is provided in the form of a family of prototypes $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_c$ and a partition matrix $U = [u_{ik}]$, $i = 1, 2, \ldots, c; k = 1, 2, \ldots, N$. The fuzzification coefficient is denoted by $m, m > 1$. Let us recall that the detailed formulas supporting computing of the partition matrix and the prototypes read as follows

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|\boldsymbol{x}_k - \boldsymbol{v}_i\|}{\|\boldsymbol{x}_k - \boldsymbol{v}_j\|} \right)^{2/(m-1)}} \tag{2}$$

$$v_i = \frac{\sum_{k=1}^{N} u_{ik}^m \boldsymbol{x}_k}{\sum_{k=1}^{N} u_{ik}^m} \quad (3)$$

where $||\cdot||$ is weighted Euclidean distance, i.e. $||\boldsymbol{x}_k - \boldsymbol{v}_i||^2 = \sum_{j=1}^{n} \frac{(x_{kj} - v_{ij})^2}{\sigma_j^2}$ with σ_j being a standard deviation of the j-th variable. The development of the partition matrix and the prototypes following the formulas listed above is realized iteratively by successively updating U and the prototypes. The key design parameter, which is of relevance in this study concerns a number of prototypes (c). The prototypes imply the number of granular descriptors of the data. The two other parameters encountered in the FCM method deal with fuzzification coefficient m and the distance function. Typically, we consider the value of the fuzzification coefficient to be set to 2, $m = 2$, and the distance is the Euclidean one or its weighted counterpart.

The prototypes can be easily interpreted by looking at their projections on the individual variables. Their projections can be ordered linearly, labeled and in this way to each prototype we assign a certain linguistic characterization as illustrated in Fig. 3. For instance, the first cluster with prototypes $\boldsymbol{v}_1$ is labeled (named) as Negative *Small* (NS) for the first variable and Negative *Small* (NS) for the second one. In other words, this cluster is a compound descriptor (Cartesian product) of well-defined semantics (Negative *Small*, Negative *Small*). Likewise the remaining clusters (information granules) can be interpreted in the same manner, say for the 4th cluster we have (Positive *Large*, Positive *Small*). In virtue of the linear order that is feasible because of the projections of the prototypes on the individual variables (making the ordering realized for a single variable), each cluster is well-interpreted.

Once the numeric prototypes have been formed, we augment their representative capabilities by spanning information granules, say intervals (hypercubes, hyper-rectangles, etc.) or fuzzy sets around them. This construct is convincing: representing numeric data exhibiting some diversity, one naturally requires that any sound representative needs to be expressed at the higher level of abstraction than the original data to capture the existing variability of the data. Hence we witness an emergence of information granules as legitimate descriptors of numeric data. Likewise, when clustering granular data (more precisely information granules of type-1), their representatives are information granules of elevated type, say type-2 information granules. It is worth noting that some studies along this line of thought have been reported in the literature. A way of forming granular prototypes was discussed in [14] while in [15] presented was an approach to building information granules in the form of shadowed sets. Some related studies are reported in [6,16].

In what follows, we develop a systematic way of constructing information granules by taking advantage of the principle of justifiable granularity established as one of the paradigms of Granular Computing.

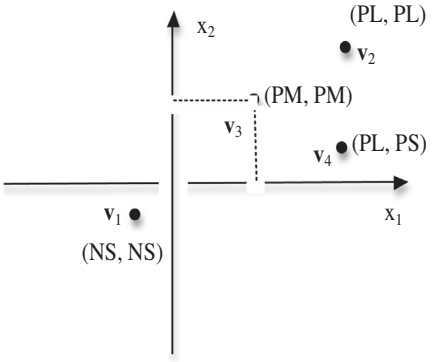

Fig. 3. Prototypes in a two-dimensional space (x_1, x_2) and their projections along with associated resulting interpretation; PL – Positive *Large*, PS – Positive *Small*, etc.

5. Formation of information granules: the principle of justifiable granularity

The principle of justifiability granularity [12] coming as one of the underlying fundamentals of Granular Computing is about forming an information granule on a basis of some experimental evidence (data). The essence of the method can be summarized as follows.

Given some one-dimensional data $\boldsymbol{X}$ and its subset $\boldsymbol{Z}$, construct an information granule G so that it satisfies two sound and intuitively appealing requirements. Let us express them in a descriptive way. First, the information granule should be experimentally justified meaning that it should "cover" enough experimental evidence. In other words, we envision the information granule to capture and be supported by the existing experimental evidence. Second, the information granule should be specific enough, which translates into a need of their easy interpretation. In this way the granule has to exhibit some tangible meaning. While this formulation of the problem has been provided in the previous studies, cf. [11,12], in what follows we develop a new algorithmic approach as we are now concerned with the *parametric* version of the principle of justifiable granularity, which requires a different algorithmic treatment of the problem.

To proceed with the details to demonstrate the optimization process behind the principle, let us start with the organization of the data that are used to construct an information granule. We consider the j-th variable ($j = 1, 2, \ldots, n$) and the i-th cluster. This leads to the data $x_{1j}, x_{2j}, \ldots x_{Nj}$ weighted the corresponding membership grades $u_{i1}, u_{i2}, \ldots, u_{iN}$.

In this study, we consider a parametric version of the principle, which involves the essentials of the data to be described. The three significant and original features of the principle considered are as follows:

(a) a parametric version of the information granule where a certain membership function is assumed in advance,
(b) incorporation of weights of data (different levels of contribution of data to the realization of information granule) that are the membership grades associated with the data.
(c) involvement of inhibitory experimental evidence (viz. data that have to be excluded for the constructed information granules).

The information granule is built around the i-th prototype in the j-th variable, namely v_{ij}. The parametric form of the information granule is described collectively by the membership function composed of two parts (segments) f and h with the condition $f(v_{ij}) = h(v_{ij}) = 1$. This fuzzy set has a finite support bounded by the points a and b. Let us start with the lower bound (a) associated with f and optimize its location by involving the principle of justifiable granularity. We calibrate the membership grades u_{ij} by considering the parametric form of f that is we introduce the calibration mechanism described as follows,

$$\phi(u_i(\boldsymbol{x})) = \begin{cases} u_i(\boldsymbol{x}) \ \textit{if} \ u_i(\boldsymbol{x}) < f(x_j) \\ f(x_j) \ \textit{if} \ u_i(\boldsymbol{x}) \geqslant f(x_j) \end{cases} \quad (4)$$

which, in essence, can be read as $\min(u_i(\boldsymbol{x}), f(x_j))$. Here the calibration pertains to the determination of the minimal value between the original membership value and the assumed parametric form of f. Refer also to Fig. 4 for the detailed notation. The criterion of coverage of data by A (more specifically by its increasing portion described by f) comes in the following form

$$f_1(a) = \max\left[0, \sum_{\substack{x \in concept \\ a \leqslant x_j \leqslant v_{ij}}} \phi(u_i(x)) - \gamma \sum_{\substack{x \notin concept \\ a \leqslant x_j \leqslant v_{ij}}} \phi(u_i(x))\right] \quad (5)$$

Note that we distinguish here between the data belonging to the concept (the first term of the above expression) and those which do not belong to the concept and as such should reduce the value of the coverage if positioned in-between a and v_{ij}. The non-negative discount factor is denoted here by γ.

Specificity [18] characterizes how detailed (specific) the constructed fuzzy set is. In general, the formal definition of specificity, $\mathrm{Sp}(\cdot)$, requires that the specificity of a single-element fuzzy set (information granule) attains its maximum, $\mathrm{Sp}(\{x\}) = 1$. Specificity exhibits a monotonicity requirement meaning that if two normal fuzzy sets satisfy the relationship $A \subset B$ then $\mathrm{Sp}(A) \geqslant \mathrm{Sp}(B)$. Furthermore $\mathrm{Sp}(\varnothing) = 0$. One the viable alternatives to define specificity is expressed as follows (here we concentrate only on the support of the fuzzy set)

$$f_2(a) = \exp\left(-\alpha \mid v_{ij} - a \mid\right) \quad (6)$$

The optimized performance index comes as a product of $f_1(a)$ and $f_2(a)$ using which we determine an optimal value of a, a_{opt}, coming as $a_{opt} = \arg\ \mathrm{Max}_a f_1(a) * f_2(a)$.

The determination of the upper bound, b_{opt} for the parametric form (h) is carried out in the same manner as presented above.

As before the strength of the inhibition component is expressed by the non-negative inhibition coefficient γ.

A construction of information granule is realized for one-dimensional data and then such information granules are aggregated in order to build an overall Cartesian product of the individual information granules. For n-dimensional data and the already constructed one-dimensional information granules $G_1, G_2, \ldots, G_n$ being formed for each variable we form the Cartesian product $\boldsymbol{G}$ with the membership function

$$\boldsymbol{G}(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n) = G_1(x_1) t\, G_2(x_2) \quad t \ldots t \quad G_n(x_n) \quad (7)$$

where t stands for any t-norm commonly used as a logic connective in fuzzy sets. For illustrative purposes we provide three example t-norms when $n = 2$

– minimum

$$\boldsymbol{G}(x_1, x_2) = \min(G_1(x_1), \quad G_2(x_2))$$

– algebraic product

$$\boldsymbol{G}(x_1, x_2) = G_1(x_1) G_2(x_2)$$

– Lukasiewicz *and* –connective

$$\boldsymbol{G}(x_1, x_2) = \max\ (0, (G_1(x_1) + G_2(x_2) - 1)$$

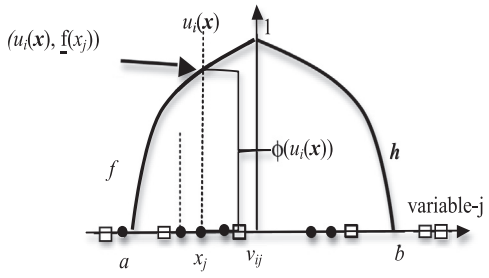

Fig. 4. Construction of the fuzzy set A with membership function described by two parts (f and h) and realized with the use of the principle of justifiable granularity.

6. Formation and interpretation of granular descriptors

The key functional components of clustering and the principle of justifiable granularity that have been discussed in the previous sections are outlined in their general format. Here in Fig. 5, we put all of them together to build a coherent development procedure and link the specific data used in the problem with the procedures discussed so far.

In light of the construction realized here, each granular descriptor $\boldsymbol{G}_i$ is described by several parameters:

linguistic characterization of the descriptor, which is possible as the prototypes projected on each variable are ordered linearly, see Fig. 3.
coverage and specificity of the granular descriptor. The coverage is computed in two steps. First, we calculate the membership function of $\boldsymbol{x}_k$ on a basis of the one dimensional membership functions G_{ij}

$$G_i(\boldsymbol{x}_k) = \min_{j=1,2,\ldots,n} G_{ij}(x_{kj}) \quad (8)$$

Second the overall coverage produced by $\boldsymbol{G}_i$ is then determined by taking the following average

$$\mathrm{coverage}_i = \frac{\sum_{k=1}^{N} G_i(x_k)}{N} \quad (9)$$

The specificity of G_{ij} (which is a fuzzy set) is determined by aggregating the specificity values for the fuzzy sets formed for individual variables. In the simplest case when information granule G_{ij} is an interval, the specificity could be expressed by taking the expression $1\text{-length}(G_{ij})/range_j$. As we are concerned with a fuzzy set (G_{ij} is described by membership function f and h), the length (G_{ij}) depends on the level of membership, viz. its value depends on the length of the β-cut of G_{ij} [11] To form the measure, which takes this dependence into consideration, we compute the following integral

$$\Phi(G_{ij}) = \int_0^1 length(\beta) d\beta \quad (10)$$

Considering the monotonicity of f and h, the bounds of the interval implied by the given threshold β are equal to $f^{-1}(\beta)$ and $h^{-1}(\beta)$; see Fig. 6.

In the sequel the specificity is computed by taking an average of the specificity results obtained for the individual variables

$$\mathrm{Specificity}\ (\boldsymbol{G}_i) = \frac{1}{N} \sum_{j=1}^{n} \left(1 - \frac{\Phi(G_{ij})}{range_j}\right) \quad (11)$$

Notably, specificity and coverage are functions of α and γ. In particular, we observe that with the increase of the values of α the coverage decreases and the specificity increases. This is the same phenomenon we have observed in case of the principle of justifiable granularity. This important relationship is displayed in Fig. 7. It sheds light on the dependencies between these two characteristics that might exhibit a visible diversity. At the same time these curves

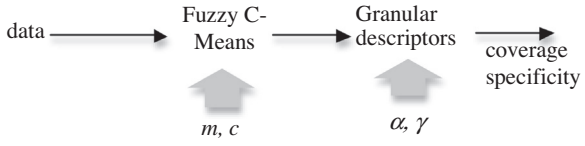

Fig. 5. A general flow of processing leading to the formation of granular data description; shown are key design parameters impact the construction of information granules.

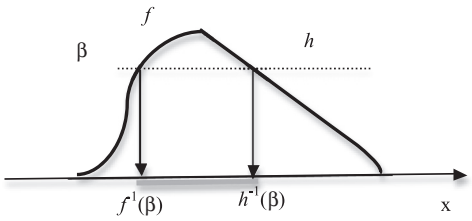

Fig. 6. Determining the length of the β-cut.

capture a holistic view at the relationships between specificity and coverage. For instance, in Fig. 7(a) as the values of a get lower, the specificity changes quite steadily (becomes lower) while the coverage increases. In contrast, in Fig. 7(b) we observe a sudden drop in the specificity when observing some increase of coverage. Likewise, the dependency visualized in Fig. 7(c) exhibits two drops in specificity when increasing the level of coverage.

To recap the overall approach from a perspective of the flow of processing, there is a critical role of the parameters associated with the essence of the method:

(a) initial buildup of the clusters (numeric prototypes). The fuzzification coefficient m and the number of clusters are the two essential parameters implying the nature of the information granules (in the sense of their resemblance of set-like information granules controlled by the values of m; the values of m close to 1 imply the Boolean-like character of the clusters). The number of clusters (c) entails a level of detail one assumes when looking at the data,
(b) the two parameters, namely α and γ, are concerned with the nature of the granular descriptors spanned over the prototypes built at the first phase of the overall process. By choosing the values of these parameters.

All in all, the four-dimensional space of parameters impacting information granules as well as their number is helpful in delivering substantial flexibility of the proposed approach and offer a designer abilities to analyze the results in a comprehensive manner.

7. Experimental studies

In this section, we report results for three data sets; two of them are from the Machine Learning repository http://archive.ics.uci.edu/ml/ and the one is coming from the PROMISE Software Engineering repository http://promise.site.uottawa.ca/SERepository/ In all experiments Fuzzy C-Means is run for selected combinations of the number of clusters (information granules) c and the fuzzification coefficient m. The clustering method was initialized randomly. The number of iterations was set to 100; it was found experimentally that this was sufficient to assure the convergence of the method; running the algorithm for the higher number of iterations did not produce any changes to the already obtained results.

The optimized function used in the construction of the granular prototypes are triangular membership functions described by (12) where the lower and upper bound are optimized through the use of the principle of justifiable granularity. Fuzzy sets used in the realization of the principle of justifiable granularity are described by triangular membership functions T assuming the following form

$$T(x;a,mod,b)=\begin{cases}\frac{b-x}{b-mod} & if\ x\ \in [mod,b]\\ \frac{x-a}{mod-a} & if\ x\ \in [a,mod]\\ 0 & otherwise\end{cases} \tag{12}$$

where a and b are the bounds of the fuzzy set while mod is the modal value of the membership function.

In the experiments, the value of g is additionally modified by the factor N/N_{neg} using which we take into account a fact that there are a few data points that are to be excluded from the data description. In other words, for the given value of ξ the corresponding value of γ is computed in the form $\gamma = \xi(N/N_{neg})$.

7.1. Synthetic two-dimensional data

The two-dimensional data are shown in Fig. 8. There are three well-delineated groups of data (black dots) for which we intend to form granular descriptors. There are also several data coming from another class (anomaly) to be excluded from the descriptors.

When running the FCM algorithm for $c = 3$ and $m = 1.1$, the obtained prototypes are as follows:

$\boldsymbol{v}_1 = [4.28\ 5.00]$, $\boldsymbol{v}_2 = [3.06\ 0.68]$, $\boldsymbol{v}_3 = [1.19\ 1.97]$. Granular prototypes $\boldsymbol{V}_1, \boldsymbol{V}_2$, and $\boldsymbol{V}_3$ spanned over the numeric prototypes

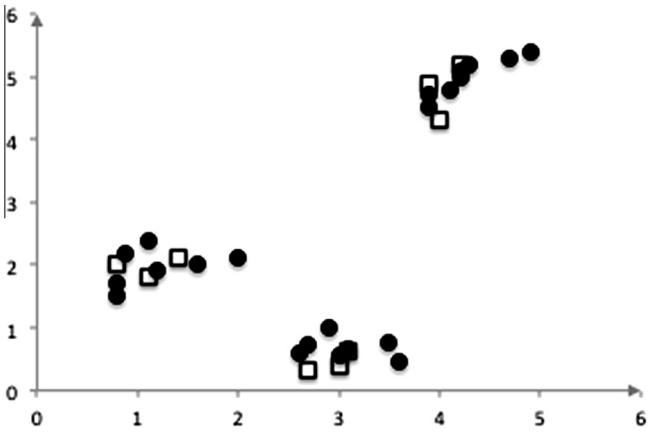

Fig. 8. Plot of synthetic two-dimensional data; the data to be excluded from the description are shown as small squares.

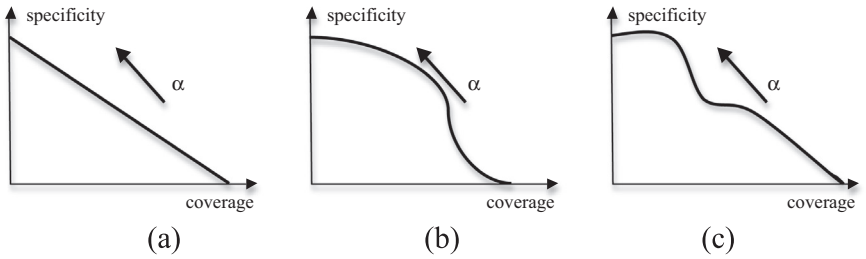

Fig. 7. Examples of relationship expressed in the specificity-coverage coordinates where using different values of α: (a) monotonic changes of the characteristics, (b) visible drop in specificity, and (c) more complicated characteristics exhibiting two points of significant changes of specificity.

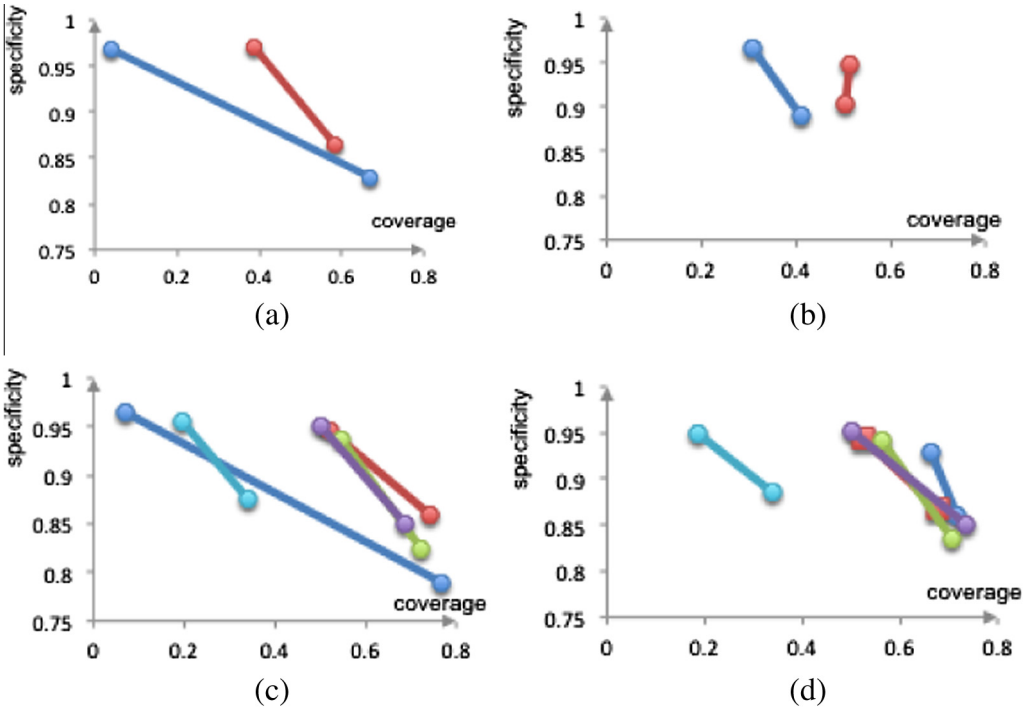

Fig. 9. Coverage – specificity plots produced for selected number of clusters and fuzzification coefficient: (a) $c = 2$, $m = 1.1$, (b) $c = 2$, m;= 2.0, (c) $c = 5$, $m = 1.1$, (d) $c = 5$, $m = 2.0$.

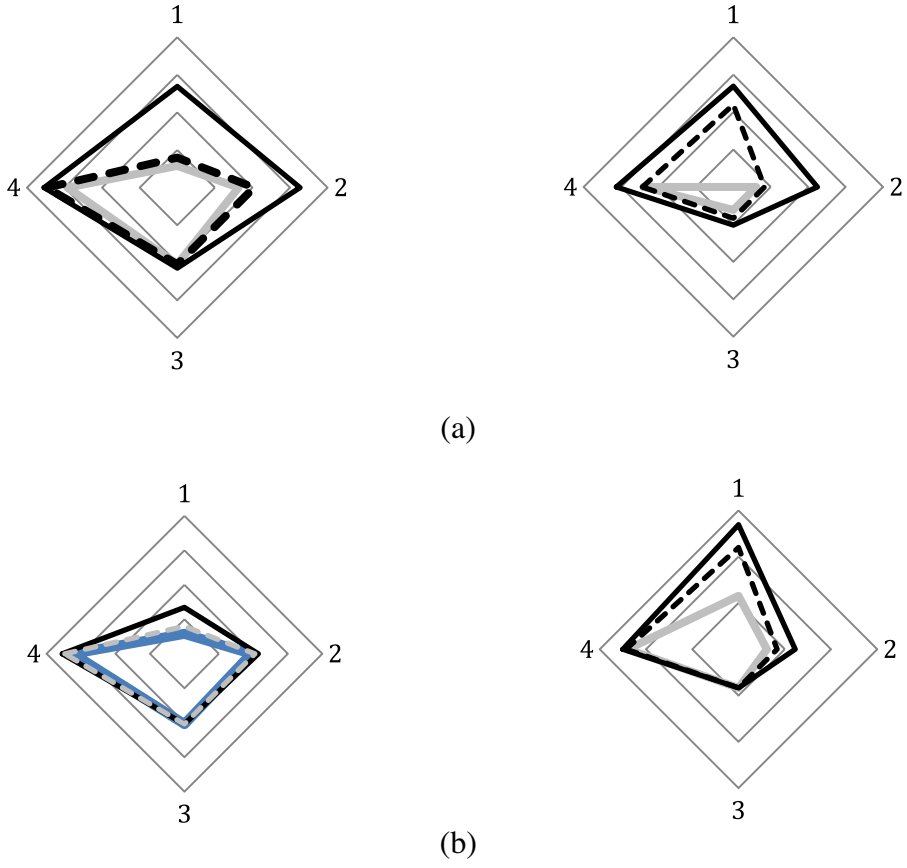

Fig. 10. Granular prototypes for $c = 2$ and selected values of parameters α and γ: (a) $\alpha = 0.05$, $\gamma = 0.0$ (b) $\alpha = 1.0$, $\gamma = 0.0$. For better visualization, the prototypes are displayed in the normalized input variables.

are formed for several combinations of the numeric values of the parameters:

$\alpha = 0.05$ $\gamma = 0.0$
$\boldsymbol{V}_1$ = [2.00 4.30] × [2.20 5.20] with the coverage and specificity equal to 0.423 and 0.147, respectively.
$\boldsymbol{V}_2$ = [0.80 4.90] × [0.60 2.40] with the coverage and specificity equal to 0.440 and 0.375, respectively.
$\boldsymbol{V}_3$ = [0.87 4.20] × [0.45 4.50] with the coverage and specificity equal to 0.400 and 0.399, respectively.

For higher value of α, $\alpha = 1.0$ $\gamma = 0.0$, one has the following granular prototypes.

$\boldsymbol{V}_1$ = [4.10 4.30] × [4.80 5.20] with the coverage and specificity equal to 0.117 and 0.926, respectively.
$\boldsymbol{V}_2$ = [2.70 3.50] × [0.60 1.00] with the coverage and specificity equal to 0.110 and 0.879, respectively.
$\boldsymbol{V}_3$ = [0.87 1.60] × [1.70 2.00] with the coverage and specificity equal to 0.176 and 0.873, respectively.

In comparison with the previous situation, it becomes apparent that the increased values of α produce more specific information granules however their coverage is reduced.
When γ assumes non-zero value, $\alpha = 0.0$ $\xi = 0.5$ one has.

$\boldsymbol{V}_1$ = [0.80 4.30] × [0.45 5.40] with the coverage and specificity equal to 0.505 and 0.346, respectively.
$\boldsymbol{V}_2$ = [0.80 4.90] × [0.60 5.40] with the coverage and specificity equal to 0.502 and 0.224, respectively.
$\boldsymbol{V}_3$ = [0.80 4.90] × [0.45 5.40] with the coverage and specificity equal to 0.462 and 0.298, respectively.

The parameters set as $\alpha = 1.0$ $\xi = 0.5$ yield the following results.

$\boldsymbol{V}_1$ = [3.60 4.30] × [4.80 5.20] with the coverage and specificity equal to 0.137 and 0.834, respectively.
$\boldsymbol{V}_2$ = [2.70 3.50] × [0.60 1.00] with the coverage and specificity equal to 0.110 and 0.879, respectively.
$\boldsymbol{V}_3$ = [0.80 1.60] × [1.70 2.00] with the coverage and specificity equal to 0.087 and 0.860, respectively.

When comparing the two combinations of the parameters α and ξ it becomes visible that the higher values of the inhibition coefficient (ξ) leads to information granules of higher specificity (and lower coverage). This is not surprising as by stressing the requirement of inhibition, the granular prototypes are naturally more confined.

7.2. Blood transfusion data set

This data set concerns 748 data coming from the donor database of Blood Transfusion Service Center in Taiwan. There are 748 donors described in the 4-dimensional input space whose features comprise: recency – months since last donation, frequency – total number of donations, monetary – total blood donated in c.c., time – months since first donation. The data are labeled with two classes describing whether the donor donated blood in March 2007 (1 stand for donating blood; 0 stands for not donating blood). There is a dominant class (0) with 76% of data belonging to this class.

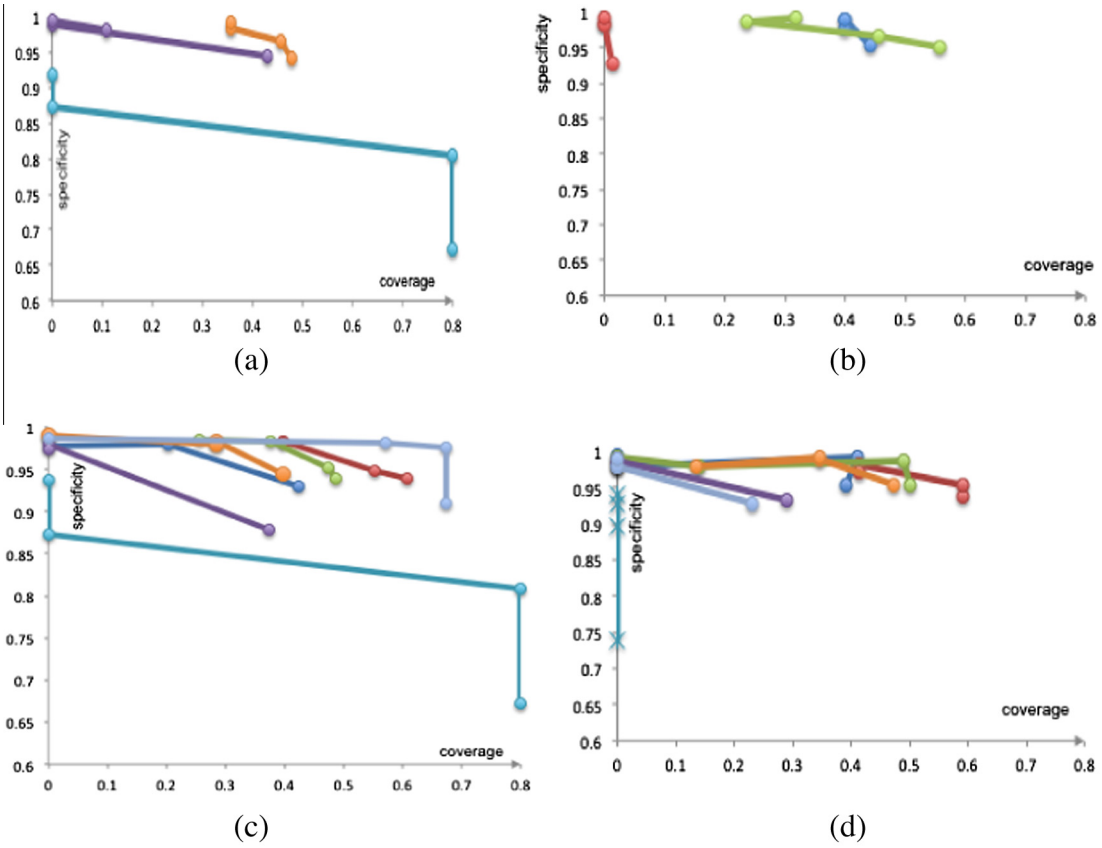

Fig. 11. Coverage – specificity plots produced for selected number of clusters and fuzzification coefficient: (a) $c = 3$, $m = 1.1$, (b) $c = 3$, $m = 2.0$, (c) $c = 7$, $m = 1.1$, (d) $c = 7$, $m = 2.0$.

The FCM algorithm is run for selected values of c and m (with the details regarding the setup of the clustering algorithm shown in the corresponding figures). The constructed coverage – specificity plots (characteristics) formed for selected values of α and γ equal to 0.0 are visualized in Fig. 9. There is a monotonic dependence between the coverage and specificity.

The prototypes can be characterized as discussed in Section 4. Here the two prototypes come with their characterization in terms

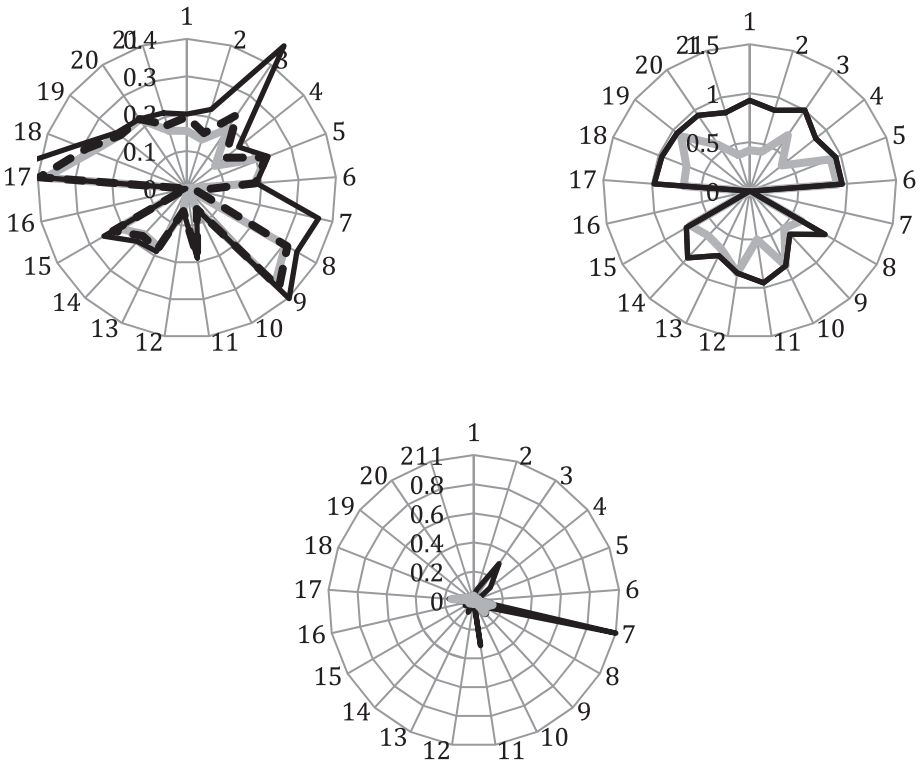

Fig. 12. Granular prototypes for $c = 3$, $m = 1.1$ and selected values of parameters α and γ : $\alpha = 0.05$, $\gamma = 0.0$. For better visualization, the prototypes are displayed in the normalized input variables.

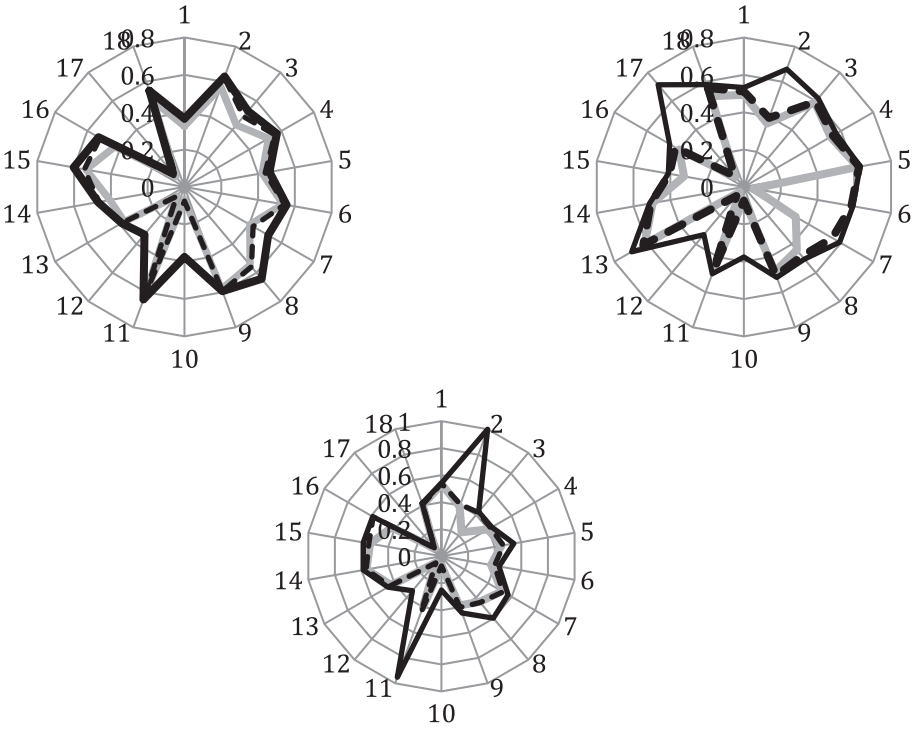

Fig. 13. Granular prototypes for $c = 3$, $m = 2.0$ and selected values of parameters α and γ : $\alpha = 0.05$, $\gamma = 0.0$. For better visualization, the prototypes are displayed in the normalized input variables.

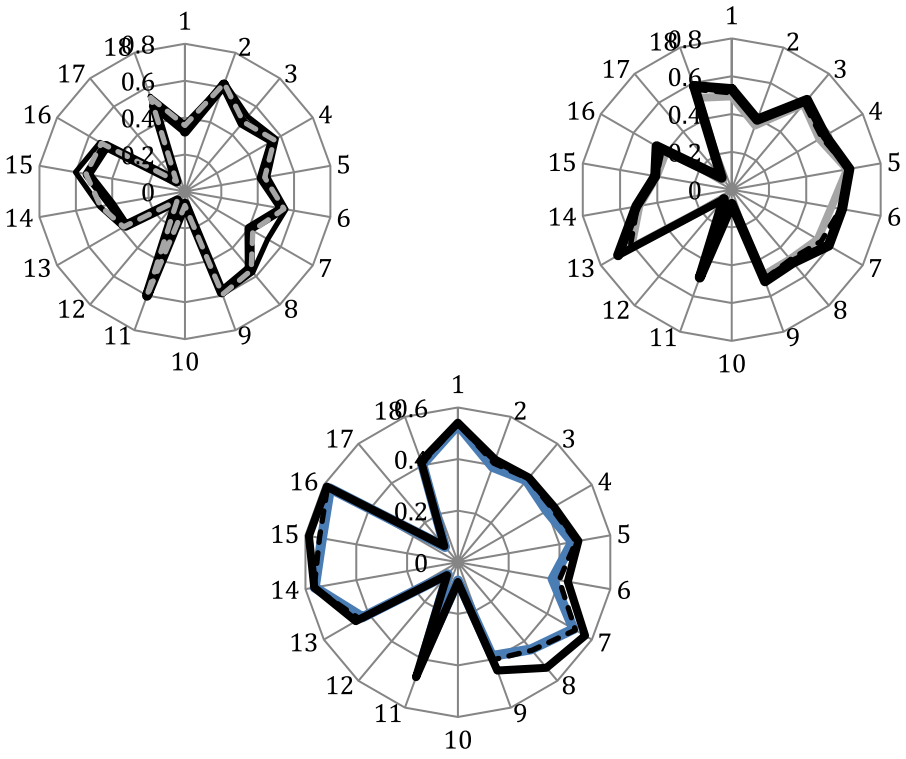

Fig. 14. Granular prototypes for $c = 3$, $m = 2.0$ and selected values of parameters α and γ : $\alpha = 1.0$, $\gamma = 0.0$. For better visualization, the prototypes are displayed in the normalized input variables.

of linguistic terms for ach variable as well as the coverage and specificity measures:

α=0.0 γ =0.0.
prototype-1 (S, S, S, S) coverage 0.67 specificity 0.82
prototype-2 (L, L, L, L) coverage 0.58 specificity 0.86

The granular prototypes built with the aid of the principle of justifiable granularity are shown in Fig. 10 in a series of radar plots. Here we present the original numeric prototypes (doted lines) along with their upper and lower (light color line) bounds. The plots visualize a location of the fuzzy sets (more specifically the position of the bounds of the support) showing how fuzzy sets are formed around the numeric prototypes.

7.3. Software data

This data set (software modules) comes from McCabe and Halstead features extractors of source code. These features were defined in the 70s in an attempt to objectively characterize code features that are associated with software quality. There are 498 data in 22-dimensional input space. These variables involve 5 different lines of code measure, 3 McCabe metrics, 4 base Halstead measures, 8 derived Halstead measures, and a branch-count. The data belong to two classes (the module has one or more reported defects or it does not have defects reported). The dominant class (90.16%) are the modules for which the defects were reported. The FCM algorithm was run for several number of clusters and selected values of m and c.

The coverage-specificity dependencies are displayed in Fig. 11 (here $\gamma = 0.0$). In most cases one can observe slight reduction in the specificity values when increasing the coverage (decreasing values of α) however there is some critical value of α when the specificity becomes significantly affected while very limited improvement of coverage is gained.

The granular prototypes are visualized in a series of plots in Fig. 12. As before the bounds of the support of the fuzzy sets of granular prototypes are visualized. The prototypes look very different in terms of their levels of granularity and the plots offer a detailed insight into the bounds of the prototypes produced for the individual variables in this way characterizing the representative regions of the input space.

7.4. Climate data

The data set describes Latin hypercube samples of 18 climate model input parameter values, predict climate model simulation crashes and determine the parameter value combinations that cause the failures. There are 540 data 18-dimensional data belonging to two classes with the dominant class composed of 494 data. The granular prototypes shown in Fig. 13 and 14 for selected values of c and m provide a detailed insight into the nature of the prototypes and a way in which their boundaries are localized around their numeric counterparts.

8. Conclusions

The study has provided a new direction in data description realized in the setting of Granular Computing. We show that information granules constitute essential descriptors of the data by delivering a concise and interpretable characterization of data. We showed that the principle of justifiable granularity supports the formation of a sound tradeoff between the legitimacy (justification) of the constructed granule and its interpretability (specificity).

There are a number of interesting direct pursuits worth considering as a follow-up of this study:

- From the optimization perspective, it could be advantageous to optimize the values of a for individual information granules. Currently there is a single value of a associated with all information granules however a tradeoff between justifiability and specificity could be set up differently for each information granule.
- This study has focused on the formation of information granules on a basis of numeric data. Naturally, a general scenario could be investigated when we are provided with granular data for which a description is to be formed. In this case one may think of a formation of information granules of so-called higher type, say being realized in the form of type-2 fuzzy sets.
- The granular description of data can serve as a starting point of further investigations along the line of system modeling, say by building granular classifiers or granular predictors. These possible avenues offer some substantial potential and help explore the directions that have not been looked at in the past. More specifically, with regard to the prediction problems one can envision a collection of granular receptive fields using which one builds a neural network with prediction results being generated by the granular descriptors in case the input is localized within one of the granular descriptors or those coming in the form of some interpolation scheme engaging granular descriptors and distances between the input and the descriptors. For the classifiers, there is a similar line of thought: the classification outcome is the one implied by a single information granule or the interpolation outcome- this may help distinguish between the quality (relevance) of the classification results.

Acknowledgments

Support from the Department for Educational Policies, Universities and Research of the Autonomous Province of Bolzano – South Tyrol is gratefully acknowledged (W. Pedrycz). Support from the Canada Research Chair (CRC) and Natural Sciences and Engineering Research Council (NSERC) is also fully acknowledged.

References

[1] F. Angiulli, Prototype-based domain description for one-class classification, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012) 1131–1144.
[2] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[3] L. Billard, E. Diday, Symbolic Data Analysis, J. Wiley, Chichester, 2006.
[4] G. Bordogna, G. Pasi, A quality driven hierarchical data divisive soft clustering for information retrieval, Knowl.-Based Syst. 26 (2012) 9–19.
[5] L. Feng, T. Li, D. Ruan, S. Gou, A vague-rough set approach for uncertain knowledge acquisition, Knowl.-Based Syst. 24 (2011) 837–843.
[6] C. Hwang, F.C. H Rhee, Uncertain fuzzy clustering: Interval Type-2 fuzzy approach to C-Means, IEEE Trans. Fuzzy Syst. 15 (12) (2007) 107–120.
[7] P. Juszczak, D.M.J. Tax, E. Pekalska, R.P.W. Duin, Minimum spanning tree based one-class classifier, Neurocomputing 72 (2009) 1859–1869.
[8] M. Kemmler, E. Rodner, E.-S. Wacker, J. Denzler, One-class classification with Gaussian processes, Pattern Recogn. 46 (2013) 3507–3518.
[9] J. Liu, Q. Hu, D. Yu, Comparative study on rough set based class imbalance learning, Knowl.-Based Syst. 21 (2008) 753–763.
[10] W. Pedrycz, Granular computing – the emerging paradigm, J. Uncertain Syst. 1 (1) (2007) 38–61.
[11] W. Pedrycz, Granular Computing: Analysis and Design of Intelligent Systems, CRC. Press/Francis Taylor, Boca Raton, 2013.
[12] W. Pedrycz, W. Homenda, Building the fundamentals of granular computing: a principle of justifiable granularity, Appl. Soft Comput. 13 (2013) 4209–4218.
[13] W. Pedrycz, Knowledge-Based Fuzzy Clustering, John Wiley, N. York, 2005.
[14] W. Pedrycz, A. Bargiela, An optimization of allocation of information granularity in the interpretation of data structures: toward granular fuzzy clustering, IEEE Trans. Syst. Man Cyber. Part B 42 (2012) 582–590.
[15] W. Pedrycz, From fuzzy sets to shadowed sets: interpretation and computing, Int. J. Intell. Syst. 24 (1) (2009) 48–61.
[16] G. Peters, Granular box regression, IEEE Trans. Fuzzy Syst. 19 (6) (2011) 1141–1152.
[17] H. Tahayori, A. Sadeghian, W. Pedrycz, Induction of shadowed sets based on the gradual grade of fuzziness, IEEE Trans. Fuzzy Syst. 21 (2013) 937–949.
[18] R.R. Yager, Ordinal measures of specificity, Int. J. General Syst. 17 (1990) 57–72.
[19] Q.Y. Yan, S.X. Xia, K.W. Feng, Probabilistic distance based abnormal pattern detection in uncertain series data, Knowl.-Based Syst. 36 (2012) 182–190.
[20] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cyber. 43 (6) (2013) 1977–1989.
[21] L.A. Zadeh, Towards a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (1997) 111–117.
[22] L.A. Zadeh, Toward a generalized theory of uncertainty (GTU)—-an outline, Inf. Sci. 172 (2005) 1–40.