Knowledge-Based Systems 101 (2016) 100–113

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

From numeric data to information granules: A design through clustering and the principle of justifiable granularity

Xin Wang[a], Witold Pedrycz[b,c,e,*], Adam Gacek[d], Xiaodong Liu[a]

[a] Department of Mathematics, Dalian Maritime University, Dalian 116026, PR China
[b] Department of Electrical & Computer Engineering, University of Alberta, 238 Elec-Comp Building, Edmonton, T6R 2V4 AB, Canada
[c] Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland
[d] Institute of Medical Technology and Equipment (ITAM), 118 Roosevelt Street, Zabrze 41-800, Poland
[e] Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University Jeddah, 21589, Saudi Arabia

ARTICLE INFO

Article history:
Received 6 November 2015
Revised 21 February 2016
Accepted 13 March 2016
Available online 17 March 2016

Keywords:
Granular computing
Granular data
Information content of information granules
Principle of justifiable granularity
Clustering

ABSTRACT

Designing information granules used intensively in Granular Computing is of paramount relevance to the fundamentals of the discipline. Information granules are key functional components in granular models, granular classifiers, and granular decision-making models. The design of information granules is central to the discipline of Granular Computing. In this study, we introduce a way of designing information granules by combining the mechanisms of unsupervised and supervised learning and subsequently using the principle of justifiable granularity. An overall design process consists of two phases. First, the granulation process involves hierarchical clustering or *K*-means clustering. It is followed by a parametric refinement of information granules realized by the principle of justifiable granularity. The characterization of information granules is offered in terms of measures of coverage, specificity, and entropy. Experimental results including synthetic data and publicly available data are covered to demonstrate the performance of the proposed approach.

1. Introduction

Granular Computing, as a coherent discipline is concerned with a construction, processing, and communicating information granules [1–3]. Information granules are fundamentals entities used in all pursuits of Granular Computing [4–9]. They are instrumental in the realization of abstraction mechanisms facilitating a way in which complex phenomena could be described, modeled, and interpreted. Information granules can be formalized in various ways and expressed in the language of set theory or interval analysis [10], fuzzy sets [11–13], shadowed sets [14], rough sets ([1,15–18], and hybrid structures such as, fuzzy rough sets, probabilistic sets. Clustering and fuzzy clustering are commonly used as an algorithmic vehicle to develop information granules [19,20]. In a nutshell, clustering transforms large collections of numeric data into a small number of information granules forming a concise and abstract description of original data.

The theory of Granular Computing comes with the principle of justifiable granularity [21]. This principle offers a certain alternative to clustering methods in realizing an important way to construct information granules on a basis of available experimental evidence and being guided by an intuitively appealing criteria of coverage of data and specificity.

In this study, we combine these two approaches and form a unique development environment of building information granules by combining mechanisms of unsupervised learning (offered by clustering techniques) and the principle of justifiable granularity augmented by the mechanisms of supervised formation of information granules. This combination can bring the advantages of these two approaches: unsupervised learning (clustering) reveals an underlying structure and brings its characterization through numeric representatives (prototypes) whereas the principle of justifiable granularity augments them by forming information granules spanned over the prototypes. More specifically, the approach advocated here consists of two main phases. First, through unsupervised learning (clustering), a collection of information granules is formed. Second, these information granules are refined with the aid of supervision mechanism incorporated within the principle of justifiable granularity. The quality of information granules is characterized with the aid of the criteria of coverage and specificity (the same as used in the principle). In addition a criterion of information content expressing homogeneity (expressed with the aid

[*] Corresponding author. Tel.: +7804398731.
E-mail addresses: xenawang7811@hotmail.com (X. Wang), wpedrycz@ualberta.ca (W. Pedrycz).

http://dx.doi.org/10.1016/j.knosys.2016.03.012
0950-7051/© 2016 Elsevier B.V. All rights reserved.

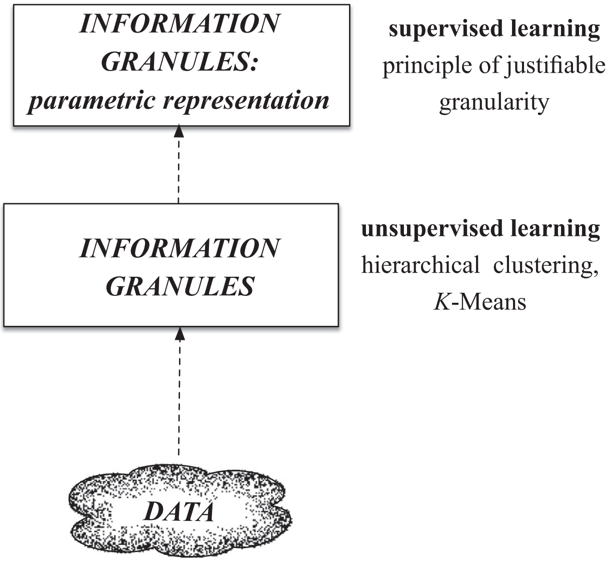
INFORMATION GRANULES: parametric representation
supervised learning
principle of justifiable granularity
INFORMATION GRANULES
unsupervised learning
hierarchical clustering, K-Means
DATA

Fig. 1. From data to granular data: a two-phase development process endowed with mechanisms of unsupervised and supervised learning.

of entropy measure) of granules formed in this way is introduced. The main interrelationships among these three descriptors of information granules are elaborated in detail and their dependence upon the values of the auxiliary parameters used in the design of information granules is established.

Let $X = \{x_1, x_2,\ldots,x_N\}$, $x_k \in R^n$ be a data set composed of elements belonging to p classes, where x_k denotes a kth data (vector of numeric entries) of dimensionality n. First, we split the data set X into c ($c \ll N$) groups (clusters) by using hierarchical clustering or clustering. This is done in unsupervised mode of learning. Second, on a basis of each cluster we develop an information granule by invoking the parametric version of the principle of justifiable granularity. At this phase, some mechanisms of supervised learning are involved. Third, we discuss a characterization of information granules constructed in this manner, which predominantly concerns their information content. The essence of the proposed scheme is captured in Fig. 1.

The structure of the paper is organized as follows. Section 2 offers a brief overview of clustering methods including the description of hierarchical clustering and K-means clustering. Section 3 gives the introduction to the parametric version of the principle of justifiable granularity. Some pertinent analysis along with illustrative examples is also covered in Section 3. In Section 4, the quality of information granules is characterized with the aid of the criteria of coverage, specificity, and information content. Experimental studies are presented in Section 6. Section 7 delivers some concluding comments.

2. Clustering methods: A brief overview

In this section, we briefly recall the essentials of clustering methods including both hierarchical clustering (as representatives of graph-oriented clustering) and objective function-based clustering coming with one of its commonly discussed representative such as K-means.

2.1. Hierarchical clustering

Clustering methods consist of separating a dataset into groups in such a way that members of the same group are more similar to one another than to the members of the other groups. Generally, clustering methods can be divided into: partitioning methods, hierarchical methods, density-based methods, graph-theoretic method, grid-based methods [22-26]. Hierarchical clustering algorithms, refer to [27], operate in one of the basic two modes. In agglomerative clustering, the process of clustering leads to the formation of clusters in a bottom-up fashion by merging the nearest clusters. In divisive clustering completed in the top-down fashion, one proceeds with splitting large clusters into smaller ones. In more detail, agglomerative and divisive strategies come in the form:

Table 1
Lance and Williams dissimilarity update formula.

Hierarchical clustering methods	Dissimilarity update formula
Single link	$\alpha_i = 0.5$, $\beta = 0$ and $\gamma = -0.5$ $(\min\{d(i, k), d(j, k)\})$
Complete link	$\alpha_i = 0.5$, $\beta = 0$ and $\gamma = 0.5$ $(\max\{d(i, k), d(j, k)\})$
Average link	$\alpha_i = \frac{\lvert i\rvert}{\lvert i\rvert+\lvert j\rvert}$, $\beta = 0$ and $\gamma = 0$
Weighted link (WPGMA)	$\alpha_i = 0.5$, $\beta = 0$ and $\gamma = 0$
Median method (Gower's, WPGMC)	$\alpha_i = 0.5$, $\beta = -0.25$ and $\gamma = 0$
Centroid (UPGMC)	$\alpha_i = \frac{\lvert i\rvert}{\lvert i\rvert+\lvert j\rvert}$, $\beta = -\frac{\lvert i\rvert\lvert j\rvert}{(\lvert i\rvert+\lvert j\rvert)^2}$ and $\gamma = 0$
Ward's method (minimum variance, sum of squared errors)	$\alpha_i = \frac{\lvert i\rvert+\lvert k\rvert}{\lvert i\rvert+\lvert j\rvert+\lvert k\rvert}$, $\beta = -\frac{\lvert k\rvert}{\lvert i\rvert+\lvert j\rvert+\lvert k\rvert}$ and γ=0

Agglomerative clustering: Each data is first assigned to a cluster containing only a single element (data). In the sequel, the pairs of clusters that are closest to each other (in sense of a certain predetermined distance) are merged into a single cluster. This merging process continues until we form a single cluster containing all the data.

Divisive clustering: Here, we proceed with a single cluster composed of all data. In the sequel, the cluster is split into two clusters where the split is guided by the separation observed between the data. The splits are carried out recursively as until the final clusters are composed of single data only.

Hierarchical clustering has been widely applied to many real-world applications areas including sensor networks, information retrieval, climate, psychology and medicine, computational biology, social sciences, and computer vision [10,28–33].

In this study, we are concerned with agglomerative clustering in order to develop information granules. A certain measure of similarity (or dissimilarity, distance) computed between any two clusters navigates the process of agglomerative clustering. As the distance is determined with regard to clusters rather than individual data, it is apparent that it could be determined in different way and this gives rise to a collection of methods referred to single linkage, complete linkage, average linkage, centroid linkage, median linkage, Ward linkage, and minimum variance linkage. In terms of dissimilarity measures used in hierarchical clustering, one can refer to the Lance-Williams dissimilarity update formula [34]. The formula points at the existence of a striking diversity of hierarchical clustering. When two data, say the ith and the jth ones are merged into a single cluster $\{i, j\}$, the dissimilarity between this cluster and other data is expressed in the following form:

$$d(\{i, j\}, k) = \alpha_i d(i, k) + \alpha_j d(j, k) + \beta d(i, j) + \gamma |d(i, k) - d(j, k)| \quad (1)$$

where α_i, α_j, β and γ are weight coefficients. For example, in the case of the single link method, one has $\alpha_i = \alpha_j = \frac{1}{2}$, $\beta = 0$ and $\gamma = -\frac{1}{2}$ which leads to the following formula (2):

$$d(\{i, j\}, k) = \frac{1}{2}d(i, k) + \frac{1}{2}d(j, k) - \frac{1}{2}|d(i, k) - d(j, k)|. \quad (2)$$

We can rewrite this expression in the following form:

$$d(\{i, j\}, k) = \min\{d(i, k), d(j, k)\}.$$