



Contents lists available at ScienceDirect

## Pattern Recognition Letters

journal homepage: [www.elsevier.com/locate/patrec](http://www.elsevier.com/locate/patrec)

## Editorial

## Data science, big data and granular mining

## Introduction

With the evolution of various modern technologies, huge amount of data is being constantly generated and collected around us. We are in the midst of what is popularly called information revolution and are living in a so-called world of knowledge. Intentionally and/or accidentally, generation of these data is inevitable. As a result, large data, broadly characterised by three Vs – large volume, velocity and variety (popularly known as “big data”) [1], is becoming a fancy word, and analysis, access and store of these data are now central to various scientific innovation, public health and welfare, public security and so on. Moreover, big data are highly complex in nature and mining them is not straight forward. Most of the information is heterogeneous, time varying, redundant, uncertain and imprecise. To reason, understand and mine useful knowledge from these data is becoming a great challenge. It is also true that large integrated data sets can potentially provide a much deeper understanding of both the nature and society, and open up new avenues for research activities.

Long before the data space, there have been histories of physical and social spaces to describe various phenomena in nature and human civilization, respectively. These two spaces eventually led to natural and social sciences, where a few research activities related to understanding of processes and concepts such as discovering various phenomena in the physical world and understanding the human interactions are being carried out. In the recent past, digitalization of various observable facts of these spaces has produced huge amount of data. With the over-flooding of data called “big data” people have started realizing the existence of data space along with the natural and social spaces, and shown remarkable interest on it to explore. Once the data are generated, they will not evolve accordingly if no special mechanism is arranged. The data may have powerful reaction to the real world even if it is fabricated, e.g., rumours spread via mobile phones or through a social media. Big data has a role in its promising usefulness in many fields such as commerce and business, biology, medicine, public administration, material science, and cognition in human brain, just to name a few. The objective of big data, as it stands, is to develop complex procedure running over large-scale enormous-sized data repositories for – extracting useful knowledge hidden therein and delivering accurate predictions of various kinds within a reasonable time period. Scientists from the academia, industry, and open source community have been trying for improved reasoning and understanding of big data and to provide better scope to solve several social and natural problems. It may be mentioned here that while analytics over big data is playing a leading role, there has been a shortage of deep analytical talent globally.

Mining of big data can be made effective with the methodologies that can deal with their characteristics, such as heterogeneity, dynamism or time varying, redundancy, uncertainty and impreciseness. Heterogeneity in data comes from the characterization of information in different ways such as using numerical, categorical, text, and image or audio/video data. Dynamism in data is due to the mechanism that generates related data changes at different times or different circumstances, which adds new uncertainty and difficulties for analysis. High dimensionality is due to the inappropriate collection of data for a task. Although a large candidate set of attributes is provided, most of them are irrelevant or redundant. These superfluous attributes deteriorate the learning performance of decision-making algorithms.

In the recent past, evolution of research interest has cropped up a relatively new area called, granular computing (GrC) [2,3,7], due to the need and challenges from various domains of applications, such as data mining, document analysis, financial gaming, organization and retrieval of huge data bases of multimedia, medical data, remote sensing, and biometrics. Several aspects of GrC [6] play important roles in the design of different decision-making models with acceptable performance.

GrC, based on technologies like fuzzy sets [9], rough sets [8], computing with words, etc., provides powerful tools for multiple granularity and multiple-view data analysis, which is of vital importance for understanding the complexity in big data. Different users may require understanding of big data at different granularity levels and need to analyse the data with different viewpoints. GrC has exhibited some capability and advantages in intelligent data analysis, pattern recognition, machine learning and uncertain reasoning for noticeable size of data. However, very few research activities have gone deep into the nutshell of heterogeneous, complex and big data analysis so far.

## Granular computing: components, characteristics and features

GrC by its name is not new, but the popularity in its use for various domains has gained recently. It is a computing paradigm of information processing that works with the process of information granulation/abstraction. GrC is an umbrella term that accommodates everything about theories, methods, techniques and tools of information granulation. In processing large-scale information, GrC plays an important role that finds simple approximate solution which is cost effective and provides improved description of real world intelligent systems. In the present scenario, synergetic integration of GrC and computational intelligence has become a hot area of research to develop and experiment various efficient decision-making models for complex problems.

Broadly, the prime motivation in the development of GrC-based methodologies is three-fold. These are: (i) GrC does not go for an excessive precision of solution which is in fact the inherent characteristic of human reasoning process, (ii) the structural re-presentation of the problem in the work-flow of GrC makes the solution process more efficient, and (iii) the computing method provides transparency in the information processing steps. Human reasoning-process normally follows the principle of information granulation and performs the computation and operations on information granules. Most often, modelling and monitoring complex systems do not require high level of precisions and in fact it is often expensive and not necessary. Further, a task with incomplete and imprecise information poses problems to differentiate distinct elements. This motivates one to go for granular representation of information and processes. As a result, GrC framework provides efficient and intelligent information processing systems for various real life decision-making applications. The said framework can be modelled with principles of neural networks, interval analysis, fuzzy sets and rough sets, both in isolation and integration, among other theories.

One of the key advantages in computing with granules is that granular processing mimics the human way of understanding of a particular problem where it requires different levels of information. GrC works in a similar fashion that deals with the structured representation of the real-world problems. In fact, this form of characterization is inherent in any task of the universe, and several relationships among the different levels of structures and multiple levels of understanding provide ample scope to understand and interpret them. In many cases, including large data sets, a problem may be difficult to be solved at one level, but subsequent derivations of the same problem on other levels may lead to a partial, or full solution. This concept of hierarchical problem abstraction is useful for knowledge discovery where the data set in question is large with many attributes. The paradigm of GrC suitably offers ample opportunities for efficient methodologies of knowledge discovery, where it can be combined with fuzzy sets and rough sets (as example); thereby developing a systematic modelling framework with a focus on the overall interpretation of the system. Such transparency in the working process provides improved interaction between the expert knowledge and the system design leading to better system performance.

The components of GrC that drive the complete process are: granules, granulation, granular relationship and computing with granules. A granule is considered as a building block, which plays a significant role in the process of GrC. Constructing appropriate granules for a particular task is crucial, because different sizes and shapes of granules decide the success rate of the GrC models. A granule is a collection of entities drawn together by indistinguishability, similarity, proximity or functionality. Significance of a granule in GrC is similar to that of any object, subset, class, or cluster in a universe. Depending on the size and shape, and with a certain level of granularity, the granules may characterise a specific aspect of the problem. Granules with different levels of granularity represent the model differently and regulate the decision-making process accordingly. In the formation of different granules, a subset of objects having similarity among them, in terms of some relation or attribute, is considered, which is a kind of clustering approach. Granules can be formed using fuzzy sets, rough sets, random set, and interval set.

The process of formation and representation of granules is called granulation. Granulation is performed in both ways, such as integrating and dispersing the granular structure. Integration involves the process of developing larger and/or higher level granules with smaller and/or lower level granules, whereas dispersion involves the process of decomposing larger and/or granules into smaller and/or lower level granules. These processes of integration and dispersion are also known as bottom-up and top-down approaches, respectively in the development of granules. Although, these two processes work

in an opposite manner, broadly they are highly correlated to each other. At higher level, a granule represents more abstract concept by ignoring irrelevant details, while at lower level it is more specific one that reveals more detailed information. A challenging problem in GrC is to coordinate and/or swap between different levels of granularity. Several attempts have been made to develop machine learning models using these approaches.

To improve the decision making-process using GrC, one may have to represent and interpret the granular relationship, such as inter and intra-relationship meticulously. The intra-relationship among the elements of a granule and inter-relationship among the elements of two different granules have strong influence in GrC-based models. These relationships provide the essential information for whether to go for integration or dispersion of granules.

It is worthwhile mentioning the importance of methods for reasoning (judgment) about properties of computations over granules. These properties are constructed (induced) using a higher level granulation over computations on granules.

### Granular computing: fuzzy and rough sets

Among the different technologies required for performing GrC, the ones based on fuzzy and rough sets are the most successful and considered to be the best choice for decision-making processes. In fact, the concept of granulation is inherent in those theories. Fuzzy set theory starts with the definition of membership function and granulates the features; thereby producing the fuzzy granulation of feature space. The fuzziness in granules and their values characterise the ways in which human concepts of granulation are formed, organised and manipulated. In fact, fuzzy information granulation does not refer to a single fuzzy granule; rather it is about a collection of fuzzy granules which result from granulating a crisp or fuzzy object. Depending on the problems and whether the granules and the process are fuzzy or crisp, one may have operations like granular fuzzy computing or fuzzy granular computing. The number of concepts formed through fuzzy granulation determines the corresponding granulation being relatively fine or coarse, and choice of the number is an application-specific optimization problem. The rough set theory provides an effective model to acquire knowledge in an information system with upper approximation and lower approximation as its core concepts and in making decisions according to the definition of indistinguishability (indiscernibility) relation and attribute reducts. Different variants of the conventional rough sets are available mainly by redefining the indistinguishability relation and approximation operators. The rough set approach (RS) can be used to granulate a set of objects into information granules (IGs). The size of the information granules is determined, e.g., by how many attributes and how many discrete values each attribute takes in the subset of the whole attribute set, which is selected to do the granulation. Generally, larger the number of attributes and more the values that each attribute takes, finer is the resulting IGs. In the perspective of knowledge transformation, the task of analysing data and solving problems by fuzzy sets or rough sets is actually to find a mapping from the information represented by the original finest-grained data to the knowledge hidden behind a set of optimised coarser and more abstract information granules. These theories are also integrated synergistically within themselves and with other knowledge acquisition models, which yield, for example, rough neural computing [5], neural fuzzy computing [11] and rough fuzzy computing [4,10]. While the former two enables forming various knowledge based granular neural networks for improved learning and performance, rough-fuzzy computing provides a stronger paradigm, than fuzzy sets or rough sets, that can handle uncertainty arising both from the overlapping characteristics of concepts/classes/regions and granularity in the domain of discourse.

Download English Version:

<https://daneshyari.com/en/article/534031>

Download Persian Version:

<https://daneshyari.com/article/534031>

[Daneshyari.com](https://daneshyari.com)