Pattern Recognition Letters 67 (2015) 109–112

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: 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.

http://dx.doi.org/10.1016/j.patrec.2015.08.001

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 irreverent 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 indistinguishibility (indiscernibility) relation and attribute reducts. Different variants of the conventional rough sets are available mainly by redefining the indistinguishibility 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.

Big data: challenges and relevance of granular mining

Various aspects of our day-to-day activities have been influenced and regularised with the presence of big data. It has not only revolutionised individuals but also affected the science, and planning and policies of the government. Although, accomplishment in this domain is in the initial stage, several technical challenges are posed that need to be addressed to fully realise the potential of big data. Generally, achieving the usefulness of big data is followed by multiple levels of operational steps, such as acquisition, information extraction and cleaning, data integration, modelling and analysis, and interpretation and deployment. Fuzzy sets, rough sets, neural networks, interval analysis and their synergistic integrations in granular computing framework have been found to be successful in most of these tasks. Research challenges around big data arise from various aspects and issues, such as their heterogeneity, inconsistency and incompleteness, timeliness, privacy, visualization and collaboration.

One may note that managing uncertainty in decision-making and prediction is very crucial for mining any kind of data, no matter small or big. While fuzzy logic (FL) is well known for modelling uncertainty arising from vague, ill-defined or overlapping concepts/regions, RS models uncertainty due to granularity (or limited discernibility) in the domain of discourse. Their effectiveness, both individually and in combination, has been established worldwide for mining audio, video, image and text patterns, as present in the big data generated by different sectors. FS and RS can be further coupled, if required, with (probabilistic) uncertainty arising from randomness in occurrence of events in order to result in a much stronger framework for handling real life ambiguous applications. In case of big-data the problem becomes more acute because of the manifold characteristics of some of the Vs, like high varieties, dynamism, streaming, time varying, variability and incompleteness. This possibly demands the judicious integration of the three aforesaid theories for efficient handling.

In this issue

This special issue on "granular mining and knowledge discovery" presents some novel approaches and methodologies reflecting the state-of-the-art of granular mining and discovering knowledge. After a rigorous review process of several submissions from all over the world, only six articles spanning a segment of the research in granular mining and its applications were selected for publication in this issue. The concepts behind the models described here would be useful to big data researchers and practitioners.

A brief description of six contributions is stated in the following in the order they appear in the issue. In their article "data granulation by the principles of uncertainty", L. Livi and A. Sadeghian present a data granulation framework that elaborates over the principles of uncertainty. It is based on the assumption that a procedure of information granulation is effective if the uncertainty conveyed by the synthesised information granule is in a monotonically increasing relationship with the uncertainty of the input data. The proposed framework is aimed to offer a guideline for the synthesis of information granules and to build the groundwork to compare and quantitatively judge over different granulation procedures. The authors have also provided suitable case studies to introduce a new data granulation technique based on the minimum sum of distances, which is designed to generate type-2 fuzzy sets. The method with automatic membership function elicitation is completely based on the dissimilarity values of the input data that makes it widely applicable.

The article "clustering in augmented space of granular constraints: a study in knowledge-based clustering" by W. Pedrycz, A. Gacek and X. Wang provides a study on fuzzy clustering in augmented space of granular constraints. The authors have described the constraints as a collection of Cartesian products of fuzzy sets. The role of information granules using order-2 fuzzy sets is underlined with regard to results of clustering produced in the transformed space. A generalization of the proposed approach with detailed algorithmic development is also discussed for the clustered data with non-numeric information.

C. Sengoz and S. Ramanna in their article "learning relational facts from the web: a tolerance rough set approach" has proposed a granular model that structures categorical noun phrase instances as well as semantically related noun phrase pairs from a given corpus. In addition they put forward a semi-supervised tolerant pattern learning algorithm that labels categorical instances as well as relations. The authors have tried to address the issue of labelling large number of unlabelled data with a small number of available labelled data. In this study, the model treats the noun phrases, which are described as sets of their co-occurring contextual patterns. For this, they have used the ontological information from the never ending language learner (Nell) system.

In the article "a new method for constructing granular neural networks based on rule extraction and extreme learning machine", the authors X. Xu, G. Wang, S. Ding, X. Jiang and Z. Zhao have introduced a framework of granular neural networks named rough rule granular extreme learning machine (RRGELM), and developed its comprehensive design process. The proposed model is based on the rough decision rules that are extracted from the training samples. In the granular neural network, the linked weights between the input neurons and granular-neurons are determined by the confidences of rough decision rules, while the linked weights between the output neurons and granular-neurons can be initialised as the contributions of the rough rules to the classification. The network is then trained with extreme learning machine algorithm and its efficacy is demonstrated on several data sets.

The next article, authored by S. Kundu and S. K. Pal, concerns with mining social networks. In their article titled "Fuzzy-rough community in social networks", a novel algorithm for community detection in a social network is described. The method identifies fuzzy-rough communities where a node can be a part of many groups with different associated membership values. The algorithm runs on a new framework of social network representation based on fuzzy granulation theory. An index, called normalised fuzzy mutual information, is defined to quantify the goodness of the detected communities. It may be mentioned here that the data available from the online social networking sites are dynamic, large scale, diverse and complex with all the characteristics of big data in terms of velocity, volume, and variety.

A. Zhu, G. Wang and Y. Dong described a robust system in "detecting natural scenes text via auto image partition, two-stage grouping and two-layer classification" to detect natural scene text according to text region appearances. Framework of this system includes three parts, such as auto image partition, two-stage grouping and two-layer classification. The first part partitions the images into unconstrained sub-images through statistical distribution of sampling points. The designed two-stage grouping method performs grouping in each sub-image in first stage and connects different partitioned image regions in second stage to group connected components (CCs) to text regions. Then a two-layer classification mechanism is designed for classifying candidate text regions.

These articles provide a wide range of methods with various characteristic features and different applications of granular computing-based mining and knowledge discovery. We hope that the publication of this issue will encourage further research activities and motivate developing novel approaches for uncertainty handling and to address the challenging issues encountered in real life problems including those in big data.

In conclusion, we would like to thank all the authors who submitted research manuscripts to this issue, as well as the anonymous reviewers for the time spent to evaluate manuscript and provide constructive comments and suggestions. We would like to extend a

special note of appreciation to Dr. Gabriella Sanniti di Baja, Editor-in-Chief of Pattern Recognition Letters, and Yanhong Zhai and Jefeery Alex, the publication supporting team for their constant support.

References

[1] E.E.S. (Eds.), Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, John Wiley and Sons, N.Y., 2015.
[2] T.Y. Lin, Granular computing: practices, theories, and future directions, in: R.A. Meyers (Ed.), Encyclopedia of Complexity and Systems Science, Springer, Heidelberg, 2009, pp. 4339–4355.
[3] W. Pedrycz, A. Skowron, V. Kreinovich (Eds.), Handbook of Granular Computing, John Wiley and Sons, London, 2008.
[4] S.K. Pal, A. Skowron (Eds.), Rough-Fuzzy Hybridization: A New Trend in Decision Making, Springer-Verlag, Singapore, 1999.
[5] S.K. Pal, L. Polkowski, A. Skowron (Eds.), Rough-neural Computing: Techniques for Computing with Words, Springer-Verlag, Berlin, 2004.
[6] S.K. Pal, S.K. Meher, Natural computing: a problem solving paradigm with granular information processing, Appl. Soft. Comput. 13 (2013) 3944–3955.
[7] J.T. Yao, A.V. Vasilakos, W. Pedrycz, Granular computing: perspectives and challenges, IEEE Trans. Cybern 43 (2013) 1977–1989.
[8] Z. Pawlak, Rough sets, Int. J. Comput. Inf. Sci. 11 (1982) 341–356.
[9] L.A. Zadeh, Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic, Fuzzy Sets Syst. 90 (1997) 111–127.
[10] P. Maji, S.K. Pal, Rough-fuzzy Pattern Recognition: Applications in Bioinformatics and Medical Imaging, Wiley-IEEE CS Press, N.Y., 2012.
[11] S.K. Pal, Granular mining and rough-fuzzy pattern recognition: a way to natural computation., (feature article). IEEE Intell. Inf. Bull. 13 (1) (2012) 3–13 (feature article).

Sankar K. Pal*
Center for Soft Computing Research, Indian Statistical Institute, 203 Barrackpore Trunk Road, Kolkata 700108, India
Saroj K. Meher
Systems Science and Informatics Unit, Indian Statistical Institute, Bangalore Centre, Bangalore 560059, India
Andrzej Skowron
Institute of Mathematics, Warsaw University, Banacha 2, 02–097 Warsaw, Poland

*Corresponding author.
E-mail addresses: sankar@isical.ac.in (S.K. Pal), saroj.meher@isibang.ac.in (S.K. Meher)