Fundamenta Informaticae 182 (2021) i-ii
DOI 10.3233/FI-2021-2067
IOS Press

Information Granulation in Data Science and Scalable Computing

Preface

This volume includes the selected extended articles which were originally presented and published as a part of the special session on *Information Granulation in Data Science and Scalable Computing*, at the IEEE International Conference on Big Data, held in Seattle on December 2018.

Information granulation, under different names, has appeared in various domains such as information hiding in programming, granularity in artificial intelligence, divide-and-conquer paradigms in theoretical computer science, fuzzy and rough set systems, neurocomputing, evolutionary algorithms, quotient spaces, belief functions, approximate computing, etc. It is associated with granular computing, which is a general approach to build efficient computational models for complex big data applications, relying on information granules such as data blocks, clusters, groups, as well as value intervals, sets, hierarchies, etc. Information granulation can be also helpful to design simplified descriptions of complex data systems and to bridge the gap between humans and intelligent systems.

Special sessions devoted to this area have been continually organized at the IEEE Big Data Conference Series, on annual basis. They have concentrated, from the viewpoint of information granulation, on important tracks such as social network computing, cloud computing, cyber-security, data mining, process mining, interpretable machine learning, knowledge management, e-intelligence, business intelligence, bioinformatics, medical informatics and IoT. In order to illustrate at least a fraction of those topics, we have invited three distinguished scientific teams to enrich their special session papers with new results, to be thoroughly evaluated by the reviewers of Fundamenta Informaticae.

The first article – titled *KNN Loss and Deep KNN*, authored by Linh Le, Ying Xie, and Vijay V. Raghavan – introduces a new approach to learning similarities between objects in the data, so as to maximize the accuracy of the k-nearest neighbors (k-NN) classification method. This methodology refers to the notion of neighborhood which is fundamental for information granulation. Instead of relying on pre-specified similarity functions, the authors define two types of k-NN loss functions whose minimization – using a supervised-style deep learning technique – leads toward formation of better similarity-based neighborhoods, delivering higher accuracy of the k-NN process.

The second article – *A Detailed Study of the Distributed Rough Set Based Locality Sensitive Hashing Feature Selection Technique*, by Zaineb Chelly Dagdia and Christine Zarges – shows that information granulation may be considered not only for objects, but also for their features, and that the

goal of grouping together similar objects or features can be not only to improve the accuracy of the data science algorithms, but also to accelerate them. In particular, the authors investigate how to use the locality sensitive hashing (LSH) technique to partition (granulate) feature spaces of big, high-dimensional tabular data sets, so as to improve the rough-set-based feature selection.

The third article – *Mining Clinical Process from Hospital Information System: A Granular Computing Approach*, by Shusaku Tsumoto, Shoji Hirano, Tomohiro Kimura, and Haruko Iwata – goes beyond typical data set formats and the corresponding decision problems, as it refers to the task of process mining, whereby first, the components of constructed processes (clinical pathways) need to be derived from a complex database as subgroupings of homogeneous cases and states. Although the authors refer to the specific application area related to hospital information systems, the proposed multi-step methodology of deriving significant pathways as the sequences of (consecutive in time) information granules is universal. This material is also a good illustration of how to combine the real data sources with the available domain knowledge, and how to take advantage of the principles of information granulation in order to build models which are interpretable for humans.

We hope that this choice of papers sufficiently emphasizes the usefulness of information granulation, and moreover, that it will inspire the readers to contribute to the future events on the corresponding topics. We would like to express our gratefulness to the authors and reviewers for their fantastic work and interaction, to the editors of Fundamenta Informaticae for their great support, to Professor Xiaohua Tony Hu who is the founder of the IEEE Big Data Conference Series, as well as to Professor Shusaku Tsumoto who has been always the main driving force behind organizing the IEEE Big Data special sessions on *Information Granulation in Data Science and Scalable Computing*.

August, 2021
Warsaw, Kaohsiung, New York

Dominik Ślęzak
University of Warsaw, Poland

Tzung-Pei Hong
National University of Kaohsiung, Taiwan

Leon S.L. Wang
National University of Kaohsiung, Taiwan