Nat Comput (2016) 15:355–357
DOI 10.1007/s11047-015-9487-4

Preface: pattern recognition and mining

Pradipta Maji · Sankar K. Pal · Andrzej Skowron

Published online: 1 February 2015

Natural computing, also called natural computation, refers to understanding computational processes observed in nature, and human-designed computing inspired by nature. It encompasses three classes of methods, namely, those that are inspired by the nature to develop novel problem-solving techniques; those that are used for modeling natural phenomena based on the use of computers; and those that employ natural materials such as molecules to compute with use of discovered models of relevant natural phenomena. Our understanding of nature, as well as the essence of computation, is enhanced if complex natural phenomena are analyzed in terms of computational processes. Characteristic for human-designed computing inspired by nature is the metaphorical use of concepts, principles, and mechanisms underlying natural systems. The processes occurring in nature can be viewed as different kinds of information processing. Self-assembly, self-reproduction, self-evolution, granulation, gene regulation networks, protein–protein interaction networks, active and passive biological transport networks, and gene assembly in unicellular organisms are some of the examples of such processes. Understanding the universe itself from the information processing point of view and engineering of semi-synthetic organisms are some efforts to understand biological systems.

The most established classical nature-inspired models of computation are cellular automata, neural computation, evolutionary computation and granular computation. More recent computational systems abstracted from natural processes include artificial life, swarm intelligence, artificial immune systems, membrane computing, DNA computing, molecular computing, quantum computing, fractal geometry, and amorphous computing, among others. In fact, all major methods and algorithms are nature-inspired meta-heuristic algorithms. Granulation is a process, among others, that is abstracted from natural phenomena. Granulation is inherent in human thinking and reasoning process. Granular computing (GrC) is a problem solving paradigm where computation and operations are performed on information granules, and it is based on the realization that precision is sometimes expensive and not very meaningful in modelling and controlling complex systems. This framework can be modeled with principles of neural networks, fuzzy sets and rough sets, both in isolation and integration, among other theories. GrC has been proven to be effective in intelligent information processing and data mining, and has a strong promise for Big data analysis.

To reflect the current trends in the domain of natural computing and its application, this special issue of Natural Computing (Springer) on *Pattern Recognition and Mining* has been brought out. The issue contains nine contributory papers, six selected from those presented in *5th International Conference on Pattern Recognition and Machine Intelligence* (*PReMI 2013*) and three out of a call for

P. Maji (✉)
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India
e-mail: pmaji@isical.ac.in

S. K. Pal
Machine Intelligence Unit, Center for Soft Computing Research, Indian Statistical Institute, Kolkata, India
e-mail: sankar@isical.ac.in

A. Skowron
Institute of Mathematics, University of Warsaw, Warsaw, Poland
e-mail: skowron@mimuw.ed.pl

A. Skowron
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

papers. The first paper entitled "*Neighborhood Granules and Rough Rule-Base in Tracking*" deals with several new methodologies and concepts in the area of rough set theoretic granular computing, which are then applied in video tracking. A new concept of neighborhood granule formation over images is introduced in this paper. These granules are of arbitrary shapes and sizes unlike other existing granulation techniques and hence more natural. The concept of rough-rule base is used for video tracking to deal with the uncertainties and incompleteness as well as to gain in computation time. A new neighborhood granular rough rule base is formulated, which proves to be effective in reducing the indiscernibility of the rule-base. This new rule-base provides more accurate results in the task of tracking. Two indices are defined to evaluate the performance of tracking. These indices do not need any ground truth information or any estimation technique like the other existing ones. All these features are demonstrated with suitable experimental results.

The differential evolution (DE) algorithm is a population based stochastic search technique widely applied in scientific and engineering fields for global optimization over real parameter space. The performance of the DE algorithm highly depends on the selection of values of the associated control parameters. Therefore, finding suitable values of control parameters is a challenging task and researchers have already proposed several adaptive and self-adaptive variants of the DE. In the second paper entitled "*Levy Distributed Parameter Control in Differential Evolution for Numerical Optimization*", the control parameters are adapted by levy distribution, named as Levy distributed DE (LdDE), which efficiently handles exploration and exploitation dilemma in the search space. In order to assure a fair comparison with existing parameter controlled DE algorithms, the proposed method has been applied on number of well-known unimodal, basic and expanded multimodal and hybrid composite benchmark optimization functions having different dimensions. The empirical study shows that the proposed LdDE algorithm exhibits an overall better performance in terms of accuracy and convergence speed compared to five prominent adaptive DE algorithms.

Outlier detection is an important data mining task with many contemporary applications. Clustering based methods for outlier detection try to identify the data objects that deviate from the normal data. However, the uncertainty regarding the cluster membership of an outlier object has to be handled appropriately during the clustering process. Additionally, carrying out the clustering process on data described using categorical attributes is challenging, due to the difficulty in defining requisite methods and measures dealing with such data. Addressing these issues, a novel algorithm for clustering categorical data aimed at outlier detection is proposed in "*Detecting Outliers in Categorical Data through Rough Clustering*" by modifying the standard *k*-modes algorithm. The uncertainty regarding the clustering process is addressed by considering a soft computing approach based on rough sets. Accordingly, the modified clustering algorithm incorporates the lower and upper approximation properties of rough sets. The efficacy of the proposed rough *k*-modes clustering algorithm for outlier detection is demonstrated using various benchmark categorical data sets.

Heuristic search is one of the fundamental problem solving techniques in artificial intelligence, which is used in general to efficiently solve computationally hard problems in various domains, especially in planning and optimization. In the paper titled "*Anytime Pack Search*", an anytime heuristic search algorithm, called Anytime Pack Search (APS), is presented, which produces good quality solutions quickly and improves upon them over time, by focusing the exploration on a limited set of most promising nodes in each iteration. The theoretical properties of APS are discussed and it has been shown that APS is complete. The complexity analysis of the proposed algorithm is also presented on a tree state-space model and it has been shown that it is asymptotically of the same order as that of A*, which is a widely applied best-first search method. Furthermore, a parallel formulation of the proposed algorithm, called Parallel Anytime Pack Search (PAPS), is presented, which is applicable for searching tree state-spaces. The completeness of PAPS is theoretically proved. Experimental results on the Sliding-tile puzzle problem, Traveling salesperson problem, and Single machine scheduling problem depict that the proposed sequential algorithm produces much better anytime performance when compared to some of the existing methods. Also, the proposed parallel formulation achieves super-linear speedups over the sequential method.

Unsupervised technique like clustering may be used for software cost estimation in situations where parametric models are difficult to develop. The paper entitled "*Software Cost Estimation Based on Modified K-Modes Clustering Algorithm*" presents a software cost estimation model based on a modified K-Modes clustering algorithm. The aims of this paper are: first, the modified K-Modes clustering which is an enhancement over the simple K-Modes algorithm using a proper dissimilarity measure for mixed data types, is presented and second, the proposed K-Modes algorithm is applied for software cost estimation. The modified K-Modes algorithm is compared with the existing algorithms on different software cost estimation datasets, and the results show the effectiveness of the proposed algorithm.

Wireless sensor network (WSN) is a special kind of ad-hoc network consisting of battery powered low cost sensor

nodes with limited computation and communication capabilities deployed densely in a target area. Clustering in WSN plays an important role because of its inherent energy saving capability and suitability for highly scalable network. The paper entitled "*A Bio Inspired and Trust Based Approach for Clustering in WSN*" presents a trust based secure and energy efficient clustering algorithm in WSN. A light weight dynamic TRUST model along with Honey Bee Mating Algorithm (HBMA) is presented, which will only prevent malicious node to be a cluster head. The choice of light weight TRUST model makes the clustering method more secure and energy efficient, which are most pivotal issues for resource constrained sensor network. A priority scheme among the trust metrics is also introduced, which is more realistic. Furthermore, the use of HBMA finds most appropriate node as cluster head. Simulation results are also presented here to compare the performance of the algorithm with Low Energy Adaptive Clustering Hierarchy and Advertisement timeout driven bee mating approach to maintain fair energy level in sensor networks.

The paper titled "*A Fast Evaluation Method for RTS Game Strategy Using Fuzzy Extreme Learning Machine*" proposes a fast learning method for fuzzy measure determination named Fuzzy Extreme Learning Machine (FELM). Moreover, it is applied to a special application domain, which is known as unit combination strategy evaluation in Real Time Strategy (RTS) game. The contribution of this paper includes three aspects. First, the authors describe feature interaction among different unit types by fuzzy theory. Second, a new set selection algorithm is developed to represent the complex relation between input and hidden layers in Extreme Learning Machine (ELM), in order to enable it to learn different fuzzy integrals. Finally, based on the set selection algorithm, the FELM model is proposed for feature interaction description, which has an extremely fast learning speed. Experimental results on artificial benchmarks and real RTS game data show the feasibility and effectiveness of the proposed method in both accuracy and efficiency.

The paper titled "*Gene Expression and Protein–Protein Interaction Data for Identification of Colon Cancer Related Genes Using f-Information Measures*" presents a new computational method to identify disease genes. It judiciously integrates the information of gene expression profiles and protein-protein interaction networks. While the *f*-information based maximum relevance-maximum significance framework is used to select differentially expressed genes as disease genes using gene expression profiles, the functional protein association network is used to study the mechanism of diseases. An important finding is that some *f*-information measures are shown to be effective for selecting relevant and significant genes from microarray data. Extensive experimental study on colorectal cancer establishes the fact that the genes identified by the integrated method have more colorectal cancer genes than the genes identified from the gene expression profiles alone. The enrichment analysis of the obtained genes reveals to be associated with some of the important KEGG pathways. All the results indicate that integrated method is quite promising and may become a useful tool for identifying disease genes.

Understanding the nature of interactions is regarded as one of the biggest challenges in projects related to complex adaptive systems. The last paper entitled "*Interactive Computations: Toward Risk Management in Interactive Intelligent Systems"* discusses foundations for interactive computations in interactive intelligent systems (IIS), developed in the Wistech program and used for modeling complex systems. The key role of risk management is emphasized in problem solving by IIS. The considerations are based on experience gained in real-life projects concerning, for example, medical diagnosis and therapy support, control of an unmanned helicopter, algorithmic trading or fire commander decision support.

As a whole, this special issue witnesses the vitality of the domain of natural computing for pattern recognition and mining. Finally, we take this occasion to especially thank Herman P. Spaink and J.N. Kok, editors-in-chief of *Natural Computing*, for giving us a new occasion to witness the vitality of natural computing as a field of computer science. We are also very thankful to B. Mitra, D. Bhattacharyya, J. Sil, J. Pal, P. Mitra, R. Murthy, S. Bandyopadhyay, S. Das, S. Paul, T. Nakashima, for their important and valuable reviews. They greatly contributed by their critical remarks and suggestions to improve the quality of the papers and the whole issue. Finally, it is our pleasure to thank the authors for their interesting and valuable contributions.