Int. J. Mach. Learn. & Cyber. (2015) 6:873–874
DOI 10.1007/s13042-015-0429-3

EDITORIAL

Editorial

Christos Anagnostopoulos[1] · Kostas Kolomvatsos[2]

Published online: 26 September 2015

Big data consist of the basis of future data driven decision making techniques. Big data domain has attracted the attention of many research institutes and companies Worldwide. The reason is that, in many applications domains, huge amount of data are produced and stored requiring the appropriate management to have the so called big data analytics. The increase on the user devices lead to an increased amount of data as well as an increased number of (multidimensional, spatio-temporal) queries over these data. The appropriate management of huge amounts of structured as well as unstructured data is the key issue for future research. Decision makers should adopt intelligent techniques over big data analytics for reaching efficient and time-optimized decisions according to the application domain. The adoption of *Machine Learning* (ML) and/or *Computational Intelligence* (CI) in handling big data could offer a number of advantages. Both ML and CI could provide means for the creation of intelligent systems that will respond to user/application queries in the minimum time together with the highest possible performance. The main focus of this special issue is on the creation/definition of intelligent techniques/systems on top of ML and/or CI methods and theories in big data and/or data streams research domains.

In response to the call of papers, we received a number of contributions in the domain of the applied ML and CI in big data settings. All manuscripts underwent a rigorous review process by three reviewers. The review process involved two or three review rounds for the majority of the manuscripts. We, finally, selected six papers to be included in this special issue.

The first paper, by Wu et al., studies a frequent pattern mining methodology. The authors deal with two challenges in order to be aligned with the new era of big data: (i) *Space complexity*: both input data, intermediate results and the outputted patterns could be too large to fit into memory which prevents many algorithms from executing; (ii) *Time complexity*: many existing approaches rely on exhaustive search or complicated data structures to mine frequent patterns which prove to be inapplicable for big data. The authors propose ISbFIM, an *Iterative Sampling based Frequent Itemset Mining* method that, instead of processing the entire data set at once, it samples computationally-manageable subsets and extracts frequent itemsets from these subsets.

The second paper by, Sidhu, P. and Bhatia, M. P. S., focuses on streaming environments and discuss an online ensemble approach, the *Diversified Online Ensembles Detection* (DOED). The DOED is capable of managing drifting concepts in large scale data streams and maintains two ensembles of weighted experts to handle the concept drift. The first ensemble exhibits low diversity while the second exhibits high diversity updated as per their accuracy in classifying the new data instances. Drifts are detected by comparing two accuracies: (i) the accuracy of an ensemble on the recent examples, and (ii) the accuracy from the beginning of the learning process. The final prediction, for an instance, is the class predicted by the ensemble which gives better accuracy.

✉ Kostas Kolomvatsos
kolomvatsos@cs.uth.gr

Christos Anagnostopoulos
christos.anagnostopoulos@glasgow.ac.uk

[1] School of Computing Science, University of Glasgow, Glasgow G12 8QQ, UK

[2] Department of Computer Science, University of Thessaly, Lamia 35100, Greece

In the next paper, a multilevel learning automata and a multilevel WalkSAT algorithm are proposed by Bouhmala, N. The proposed models are adopted as a paradigm for finding a tactical interplay between diversification and intensification for large scale optimization problems. The multilevel paradigm involves recursive coarsening to create a hierarchy of increasingly smaller and coarser versions of the original problem. The process is repeated till the size of the smallest problem falls below a specified threshold. A solution for the problem at the coarsest level is generated, and then successively projected back onto each of the intermediate levels in reverse order. The solution, at each child level, is improved before moving to the parent level.

The next paper, by Ludwig, S. A., studies the parallelization and scalability of the Fuzzy C-Means clustering algorithm. The algorithm is parallelized using the MapReduce paradigm outlining how the Map and Reduce primitives are implemented. A validity analysis is conducted in order to show that the implementation works correctly achieving competitive purity results compared to state-of-the-art clustering algorithms. In addition, a scalability analysis is conducted to demonstrate the performance of the parallel Fuzzy C-Means implementation with increasing number of computing nodes.

Gambhir, D. and Rajpal, N. present the pairFuzzy algorithm responsible to produce a high visual quality image at low bit rates when adopting large scale image data. The described algorithm is simple and efficient compared to JPEG. The proposed algorithm is carried out in three steps. First, an image is preprocessed using competitive fuzzy edge detection which efficiently detects the edge pixels contained in the image. Second, based on the edge information the image is compressed and decompressed using improved fuzzy transform. Third, the reconstructed image is post processed using fuzzy switched median filter for artifact reduction.

Finally, in the last paper, Boratto, L. and Carta, S. present a set of group recommender systems that automatically detect groups of users by clustering them, in order to respect a constraint on the maximum number of recommendation lists that can be produced. Group recommendations are useful when the number of recommendation lists that can be generated is limited. In such a case, grouping users and producing recommendations to groups becomes necessary, especially when large datasets containing users-related data are present. The proposed systems have been largely evaluated on two real-world datasets and validated with hundreds of experiments and statistical tests, in order to compare them.

All together, we hope that this special issue will provide new inputs for those working in the big data applications, as well as for those who are interested in new developments of big data analytics in general.

The Guest Editors
Dr. Anagnostopoulos, Christos
Dr. Kolomvatsos, Kostas