Intelligent Data Analysis 23 (2019) 1–3
DOI 10.3233/IDA-180895
IOS Press

Editorial

Dear Colleague:
Welcome to volume 23(1) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, is the first issue for our 23rd year of publication. It contains 12 articles representing a wide range of topics related to the theoretical and applied research in the field of Intelligent Data Analysis.

The first two articles are about various forms of data preprocessing in IDA. Qiu and Xiang in the first article of this issue explain feature selection in large data analysis applications and explain that Particle Swarm Optimization (PSO) was originally designed for continuous optimization problems and the discretization in feature selection. The authors introduce a novel feature selection algorithm based on a set based discrete PSO (SPSO) which employs a set based encoding scheme which makes it able to characterize the discrete search space in feature selection problem. The authors also introduce a novel feature subset evaluation criterion based on contribution rate. The proposed methods are compared with six filter approaches and four wrapper approaches on ten well known UCI dataset. Poretela *et al.* in the second article of this group discuss the overall topic of outliers and their detection and argue that outliers are affected by the so-called Simpson paradox: a trend that appears in different groups of data but disappears or reverses when these groups are combined. The authors propose that one can learn a regression tree that could grow by partitioning the data into groups more and more homogeneous of the target variable. They authors observe that some points previously signaled as outliers are no more signaled as such, but new outliers appear.

The second group of articles in this issue are about various forms of learning in IDA. Li *et al.* in the third article of this issue discuss the advantages of integrating expert knowledge for Bayesian network structure. The authors propose a new structure learning method to improve the performance of expert knowledge usage and the intuitionistic fuzzy set to express and integrate the expert knowledge. Their experiments demonstrate the validity of the proposed scheme and their comparison with the existing research results. Djenouri *et al.* in the next article of this issue argue that Association Rule Mining (ARM) is time-consuming on big datasets and there is a need for developing new scalable algorithms for this class of tasks. The authors propose to design parallel algorithms using the massively parallel threads of a GPU processor. The idea is to improve the existing approaches by parallelizing the other steps of the Bee Swarm Optimization process where the authors introduce three new algorithms. Their experimental results show that their proposed approach outperforms the three other approaches in terms of processing speed. Mortazavi-Dehkordi and Zamanifar in the next article of this issue discuss the topic of efficient resource scheduling for the analysis of big data streams. The authors present an efficient resource scheduling framework, used by streaming Big Data analysis applications based on cluster resources. This framework proposes a query model using Directed Graphs (DGs) and introduces operator assignment and operator scheduling algorithms based on a novel partitioning algorithm. Their experiments with the benchmark and well-known real-world queries show that their proposed idea can significantly reduce the latency of streaming Big Data analysis. Ray *et al.* in the sixth article of this issue propose an efficient, approximate algorithm to solve the problem of finding frequent subgraphs in large streaming graphs where the graph stream is treated as batches of labelled nodes and edges. The computational complexity of their proposed approach is bounded to linear limits by looking only at the changes made by the most recent batch, and the historical set of frequent subgraphs. The proposed approach also includes a sampling algorithm that samples regions of the graph that have been changed by the most recent update to the graph. The performance of the proposed approach is evaluated using five large graph datasets, where it is shown to be faster than the state of the art large graph miners while maintaining their accuracy.

The third group of articles are about the topic of sentiment analysis. Ahmad *et al.* in the first article of this group argue that in sentiment analysis, the high dimensionality of the feature vector is a key problem because it can decrease the accuracy of sentiment classification and make it difficult to obtain the optimum subset of features. To solve this problem, the authors propose a new text feature selection method that uses a wrapper approach, integrated with ant colony optimization (ACO) to guide the feature selection process. The performance of their proposed algorithm on customer review datasets is evaluated and compared with two hybrid algorithms from the literature, namely, the genetic algorithm with information gain and rough set attribute reduction. The results of their experiments showed that the proposed algorithm was able to obtain the optimum subset of features and can improve the accuracy of sentiment classification. In a similar study from the same research group, Ahmad *et al.* present an in-depth review of feature selection techniques in sentiment analysis (SA). The authors offer an overview of the role and techniques of feature selection (FS), Sentiment Words (SWs) detection, and the identification of the relationship between features and SWs. The main contributions of this review are its sophisticated categorisations of a large number of recent articles related to FS techniques and the detection of SWs. This review also looks at the metaheuristic approach as a FS technique in SA. Along the same line of research, Keith Norambuena *et al.* in the last article of this group apply sentiment analysis to opinion mining of scientific paper reviews. The main objective of this analysis was to automatically determine the orientation of a review and contrast it with the assessment made by the reviewer of a given article. This would allow scientists to characterize and compare reviews crosswise and more objectively support the overall assessment of a scientific article. Their results include a set of experiments conducted to evaluate the capability and performance of the proposed approaches relative to a baseline. Their results show improvements in the case of binary, ternary and a 5-point scale classification in relation to classical machine learning algorithms such as SVM and Naive Bayes.

And finally, the last group of articles in this issue are about enabling techniques and novel methods. The first article of this group by Joutsijoki is about classification of patients and controls based on stabilogram signal data, focusing on inner ear balance problem. The authors apply a wide variety of machine learning algorithms from traditional baseline methods to state-of-the-art techniques such as Least-Squares Support Vector Machines and Random Forests. Their results show that machine learning algorithms are well capable of separating patients and controls from each other. Vu *et al.* in the second article of this group present a density-based clustering applied to facial expression recognition. The authors propose a novel semi-supervised density based clustering which integrates effectively several kinds of side information, and embeds an active learning strategy in order to find the most valuable clusters. The authors present a series of experiments on both synthetic and real world data sets. The last article of this issue by Alarcón-Paredes *et al.* is about a novel application on gene selection for enhanced classification of microarray data. The authors propose to use a feature ranking and weighting scheme, which combines statistical techniques with a weighted k-NN classifier using a modified forward selection procedure. They compare their approach to state-of-the-art feature selection algorithms by means of the Friedman test. Their experimental results show the classification superiority of their method on most of the gene expression datasets used in their study.

In conclusion, we would like to thank all the authors who have submitted the results of their excellent research to be evaluated by our referees and published in the IDA journal. We look forward to receiving your feedback along with more and more quality articles in both applied and theoretical research related to the field of IDA.

With our best wishes,
Dr. A. Famili
Editor-in-Chief