Intelligent Data Analysis 20 (2016) 975–977
DOI 10.3233/IDA-160852
IOS Press

Editorial

Dear Colleague:
Welcome to volume 20(5) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, the fifth issue of 2016, consists of twelve articles, all covering a wide range of topics related to the theoretical and applied research in the field of Intelligent Data Analysis.

The first three articles of this issue are about various aspects of data preprocessing in IDA. Zhang *et al.* in the first article of this issue discuss environmental sensing using multitudes of wirelessly connected sensors where commodity sensors are installed and the data are known to be unreliable and noisy. The authors propose a sensor reliability-based cleaning method, called Inuence Mean (IM), which uses weighted aggregation based on individual sensor reliabilities. Their experimental results show that the method generally improves the data cleaning accuracy, particularly when the behaviors of unreliable sensors vary drastically from reliable sensors. Nemtsov *et al.* in the second article of this issue discuss matrix compression using the Nystrom method where they extend the applicability of this method and describe how it can be applied to find the singular value decomposition (SVD) of general matrices and the eigenvalue decomposition (EVD) of square matrices. There are three contributions to their proposed method: first, it allows the compression of a general matrix M; second, it allows the approximation of the SVD and EVD when they cannot be directly calculated due to space and time limitations; and third, a novel algorithm for selecting the initial sample is presented. Salama *et al.* in the third article of this group discuss the topic of data reduction for classification using ant colony algorithms, and present a new ant colony optimization (ACO) algorithm for data reduction which is based on both feature and instance selection. The approach is intended to improve the predictive quality of the constructed classification models. Their empirical evaluations on a large number of benchmark datasets with five well-known classification algorithms show that their approach improves the predictive quality of the produced classifiers.

The next group of articles are about various forms on learning. Vagner, in the first article of this group, introduces the OPTICS algorithm which is a hierarchical density-based clustering method. The proposed algorithm builds a grid structure to reduce the number of data points and in order to get the clusters, the algorithm uses the reachability plots of the grid structure to determine to which cluster the original input points belong. The experimental results presented in the paper show that the proposed approach is fast and the speed-up can be one or two orders of magnitude or more, depending mainly on certain parameter of the algorithm. Wang *et al.* in the fifth article of this issue, discuss ensemble learning via manipulating the training set and present a strategy which is intended to combine learning set resampling and random subspace method applied to high-dimensional domains. The authors propose a new procedure, called Bag of Little Bootstraps on Features (BLBF), which works by combining the results of bootstrapping multiple feature subsets of the original dataset using the random subspace method. Their empirical experiments on various high-dimensional datasets demonstrate that the proposed approach outperforms the state-of-the-art instance-based resampling learning algorithms. Ferber-Hernandez *et al.* in the next article propose a Sequential-Patterns Classifier, which is based on a novel pruning strategy, using the Netconf as measure of interest, that allows to prune the rules search space for building specific rules with high Netconf. The proposed classifier was evaluated using Weka, a popular suite of machine learning software. The experiments reported were conducted using several document collections. Mehrizi and Sadoghi-Yazdi in the seventh article of this issue present an analytical semi-supervised learning method that is based on Growing Self Organized Maps (GSOM) and extreme learning machine. Extreme learning machine is used in this research to exploit the substantial classification response and the learning of GSOM parameters are eliminated with use of the extreme learning machine. The proposed method has been applied on the online and partially labeled dataset where it is reported that the F-measure of proposed method is more precise than the conventional semi-supervised GSOM. Guerine *et al.* in the last article of this group discuss integrating data mining techniques with metaheuristics in order to obtain patterns of suboptimal solutions that are used to guide the heuristic search for better-cost solutions in less computational time. One of the challenges of this work would be to extend this hybrid approach to a broader domain. Therefore the authors propose a hybrid data mining heuristic to solve the one-commodity pickup-and-delivery traveling salesman problem, for which solutions are defined by sequences of elements. Computational experiments, reported in this article on a set of instances from the literature, showed that the hybrid heuristic would reach better-costs solutions faster than other strategies.

The third group of articles in this issue are on applied IDA research where each article introduces a novel approach. Kim and Yun in the first article of this group discuss mining high utility itemsets based on a time decaying model and argue that most of the existing high utility itemset mining methods cannot efficiently work in terms of both runtime and memory usage. The authors propose a new tree-based algorithm that mines recent high utility itemsets over data streams. On the basis of the time decaying model, the proposed algorithm diminishes the utilities of transactions according to their arrival-time in order to assign larger weights to recent data compared to those of older ones. Experimental results presented in this article demonstrate that the proposed algorithm can mine recent high utility itemsets from varying stream of data while consuming smaller computational resources than those of the existing ones. Shakya *et al.* in the second article of this group argue that micro-level (atomic-scale) activities are considered to be the key to understanding various macro-level (bulk) properties of a material. It is further argued that mining the structure graph of any material could offer little help in this scenario and it is the patterns among the atomic dynamics that may reveal the mechanisms underlying a particular material property. Therefore discovery of such patterns can lead to a better model and better predictions of the properties and their behaviors. The authors propose an event graph to model the atomic dynamics and a graph mining algorithm to discover popular subgraphs in the event graph. Their experiments with simulation data of silica liquid demonstrate the effectiveness of their mining system. Ben Ishak and Feki in the eleventh article of this issue introduce a novel application that discusses the problem of knowledge extraction within the banking domain using statistical learning systems. The authors perform an intensive comparative study on various banking systems so that they can propose a variable ranking and selection within a nonlinear multiclass framework. Their experiments that are performed on different simulated datasets and a real dataset show that Random Forest (RF) is slightly better than Support Vector Machine. The authors also show that RF is more robust to the selection bias problem and classification accuracy is slightly improved by the ratios selection. And finally Shen and Wang in the last article of this issue discuss smart grid systems that are becoming common in many countries, and how the power consumption data gathered from smart meters could allow electricity companies to better understand electricity usage in the future and monitor electricity supply more efficiently. The authors propose an electricity consumption forecasting model established on the framework of support vector regression (SVR). The study involves various consumption patterns where the stepwise regression analysis is applied for feature selection. The accuracy of the SVR model developed in this study is largely dependent on the selection of the model parameters. The particle swarm optimization (PSO) algorithm is also proposed to determine the optimal values of parameters that improve the accuracy and efficiency of the SVR model. Their experimental results show that the proposed method can provide electricity forecasts with 0.70% and 2.55% of mean absolute percentage error for 1 and 24 hours ahead, respectively.

In conclusion, with this issue of the IDA journal, we would like to let you know that the IOS Press, the publisher of the IDA journal, is celebrating the 20th anniversary of the IDA journal where they have special events during ECML/PKDD 2016 in Verona-Italy. This year, in addition to our six regular issues, we also published a special issue related to CIARP series of conferences. 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

e