Intelligent Data Analysis 21 (2017) 755–757
DOI 10.3233/IDA-170886
IOS Press

Editorial

Dear Colleague:
Welcome to volume 21(4) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal that 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 three articles of this issue are about various aspects of data preprocessing. In the first article, Li and van der Aalst discuss the effects of deviations in complex event logs that are often negative, but sometimes also positive and argue that it is useful to detect deviations from event logs which record various behaviors of the organization. The authors propose a novel approach that is faster than cluster-based methods and is less time-consuming than creating clusters. The experiments presented here show that the proposed approach is also more accurate than model-based approaches. Akiko *et al.* in the second article of this issue argue that benchmarking is among the most widely adopted practices in business and emphasize that conducting multidimensional benchmarking in data warehouses has not been explored from a technical efficiency perspective. The authors formulate benchmark queries in the context of data warehousing and business intelligence, and develop algorithms to answer benchmark queries efficiently. Their empirical study using the TPC-H and the weather data sets demonstrates the efficiency and scalability of their proposed methods. Aldana-Bobaldialla *et al.* in the third article of this issue emphasize that a common task in data analysis is to find the appropriate data sample whose properties allow us to infer the parameters and behavior of the data population. The authors also explain that the effectiveness of such ways is bounded by several considerations such as the sampling strategy, the size of the population and the dimensionality of the space of the data. The authors propose a method based on a measure of information in terms of Shannon's Entropy and introduce a breed of Genetic Algorithms called Eclectic Genetic Algorithm. They evaluate the proposed method with synthetic datasets; where their results show that the proposed method is highly suitable based on its effectiveness to visualize data reduction in different applications.

The next six articles are on various forms of supervised and unsupervised learning. Khadidja *et al.* in the first article of this group argue that conventional clustering algorithms optimize a single criterion, which may not conform to diverse needs of multidimensional data science. The authors propose a new clustering algorithm that solves multiple clustering issues and it simulates a proposed Marked Point Process to find clusters of complex shapes present in the raw data space. The results of the proposed algorithm proves its efficiency on high complex and scalable datasets. Zhou *et al.* in the fifth article of this issue discuss that social media provides unprecedented opportunities for people to disseminate information and share their opinions and views online. They also emphasize that extracting events from social media platforms such as Twitter could help in understanding what is being discussed. However, event extraction from social text streams poses huge challenges due to the noisy nature of social media posts and dynamic evolution of language. The authors propose a generic unsupervised framework for exploring events on Twitter which consists of four major steps, filtering, pre-processing, extraction and categorization, and post-processing. Their results show high precision is achieved for event extraction using their Bayesian model, outperforming a competitive baseline substantially. Khashei in the sixth article of this issue emphasize that in clustering, the most important and widely used technique for data

1088-467X/17/$35.00 © 2017 – IOS Press and the authors. All rights reserved

exploration and knowledge discovery, obtaining clusters that exhibit within-cluster high similarity or homogeneity and between-cluster high dissimilarity or heterogeneity, depends on the similarity notion, which has not been clearly defined for clustering purposes. The authors proposed the learning speed of the supervised neural networks as novel intelligent similarity measurement for unsupervised clustering problems. Their empirical results of the simulated data sets indicate that their proposed method not only can be used as similarity measurement in clustering tasks, but also can produce highly accurate results. Liu *et al.* in the next article of this group argue that multi-label learning has attracted significant attention from machine learning and data mining and although many multi-label classification algorithms have been devised, few research studies focus on multi-assignment clustering. The authors propose a nonparametric multi-assignment clustering (MAC) algorithm, which allows the model complexity to grow as more data instances are observed. The proposed algorithm determines the number of clusters from data, so it provides a practical model to process massive data sets. The article includes an evaluation metric for MAC based on the characteristics of clustering and multi-assignment problems. Their experiments on two real data sets indicate that the proposed method is competitive and outperforms the alternatives on most data sets. Salama *et al.* in the eighth article of this issue discuss Instance-Based Learning methods that predict the class label of a new instance based directly on the distance between the new unlabeled instance and each labeled instance, without constructing a classification model. The authors introduce a novel class-based feature weighting technique, in the context of instance-based distance methods, and present three different approaches of instance-based classification: k-Nearest Neighbours, distance-based Nearest Neighbours, and Gaussian Kernel Estimator. The authors empirically evaluate the performance of their proposed algorithms on 36 benchmark datasets, and compare them with conventional instance-based classification algorithms, using various parameter settings, as well as with a state-of-the-art co-evolutionary algorithm for instance selection and feature weighting for Nearest Neighbours classifiers. Maldonado *et al.* in the last article of this issue introduce an intelligent system for modeling the student enrollment decisions problem. The approach is based on a nested logit classifier to predict which prospective students will eventually enroll in different Bachelor degree programs of a small-sized, private Chilean university. The authors perform feature selection to identify the key features that influence the student decisions, such as socio-demographic variables (gender, age, school type, among others), admission efforts, and admission test results. The results presented in the article suggest that on-campus activities are far more productive than career fairs and other efforts performed off campus, demonstrating the importance of bringing prospective students to the university.

The last three articles in this issue are about novel applications in IDA. Keyvanrad and Homayounpour in the first article of this group discuss deep belief networks that can extract suitable features from large data sets where one of the important improvements is sparsity in hidden units. They argue that one of the main problems in sparsity techniques is to find the best hyper-parameter values which need dozens of experiments to obtain. The authors propose a dynamic hyper-parameter value setting for resolving this problem. According to their results, their proposed dynamic method achieves acceptable recognition accuracy on test sets in different applications, including image, speech and text. Rauch and Simunek in the next article of this group compare two association rules data mining methods called the *apriori* and ASSOC procedures. An association rule is understood as an implication between conjunctions of attribute-value pairs. The ASSOC procedure mentioned in this article is an implementation of the GUHA method of mechanizing hypothesis formation developed since the 1960s. Arules is a computational environment for mining association rules based on *apriori* and the 4ft-Miner procedure The authors show that the Arules approach to missing information can lead to a large number of misleading rules. It is also shown that a secured completion developed for the ASSOC procedure avoids this problem.

And finally, Dibiaso Rossi *et al.* in the last article of this issue discuss the problem of selecting learning algorithms in data mining and an important task for the success of a meta-learning system which is gathering data about the learning process. Majority of data mining systems are built under the assumption that the data are generated by a stationary distribution, i.e., a learning algorithm will perform similarly for new data from the same problem. However, many applications generate data whose characteristics can change over time. Therefore, a suitable bias at a given time may become inappropriate at another time. The authors provide a set of guidelines to support the proposal to describe non-stationary data over time. Their experimental results using real data streams showed the effectiveness of the proposed data characterization general scheme to support algorithm selection by meta-learning systems.

In conclusion, we would like to thank all the authors who have submitted the results of their excellent research to be published in this issue of 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. Famil
Editor-in-Chief