Intelligent Data Analysis 19 (2015) 949–951
DOI 10.3233/IDA-150752
IOS Press

Editorial

Dear Colleague:

Welcome to volume 19(5) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, the fifth for 2015, contains twelve articles, all covering a wide range of applied and theoretical research related to the field of Intelligent Data Analysis.

The first five articles are on various forms of learning and classification. Gordon *et al.* in the first article of this issue explain shapelets as a novel form of classifying time-series data and argue that the disadvantage of current shapelet methods is their high time and memory consumption. A shapelet is defined as a subsequence extracted from one of the time-series in the data set which best separates between time-series coming from different classes of the data set. The authors introduce an algorithm for randomized model generation for shapelet-based classification that can generate a model with surprisingly high accuracy after evaluating a small fraction of the shapelet space. The approach is evaluated using a large number of data sets. Yuan *et al.* in the second article of this issue argue that in order to discover rules in the context of time-series data, a symbolic aggregate approximation (SAX) representation could be applied to discretize the real-valued and high-dimensional time-series data into segments and convert each segment to a symbol. On this basis, the authors propose an algorithm to discover Class Sequential Rules (CSRs) and make the final prediction at first. They then apply a new lazy associative classification method, in which the computation is performed on a demand driven basis. Various experimental results presented in this article show that the lazy associative classification for time-series can be interpretable and competitive with the current state-of-the-art algorithm. In the third article of this issue, Amirkhani and Rahmati argue that some of the basic algorithms for learning the structure of Bayesian networks, such as the well-known K2 algorithm, require a prior ordering over the nodes as part of their input. As an alternative, the authors introduce the aggregation of ordering information provided by multiple experts to obtain a more robust node ordering. In their proposed approach, the accuracies of participants, not known in advance, are estimated by the expectation maximization algorithm. Their experimental results demonstrate the effectiveness of the proposed method in improving the structure of a given learning process. Loterman and Mues, in the fourth article of this issue, discuss the complexity of selecting the most suitable algorithm for a learning task and discuss the idea of meta-learning where they explore to what degree dataset characteristics can help identify which regression/estimation algorithm will best fit a given dataset. The authors focus on comprehensible 'white-box' techniques in particular (i.e. linear, spline, tree, linear tree or spline tree) as those are of particular interest in many real-life estimation settings. The authors found that algorithm based characteristics such as sampling landmarks are major drivers for successfully selecting the most accurate algorithm. In the last article of this group, Su *et al.* discuss rule-learning that is applied to extract knowledge from a dataset and represent it in a form that is easy for people to understand. They explain two popular forms of rule learning, RIPPER and PART and argue that one has an overprunning problem and the other skew sensitivity. To overcome this problem, they propose a K-L divergence-based method that uses K-L divergence as a splitting criterion to build partial decision trees. They present a wide range of experiments on imbalanced datasets in combination with SMOTE processing. The results obtained, which contrasted through nonparametric statistical

tests, show that their proposed approach is robust in the presence of class imbalance, especially when combined with SMOTE.

The next group of articles in this issue are on the topic of learning and optimization. Tran *et al.* present an improved approach of particle swarm optimization (POS) that is based on a new neighborhood search strategy with diversity mechanism and Cauchy mutation operator. The authors evaluate their proposed approach on a number of well-known benchmark functions, where they demonstrate that the proposed algorithm has significant improvement over several other PSO variants for global numerical optimization. Part of their experimentation is on data clustering where their experimental results on artificial and real-world data sets show their proposed method outperforms other comparative clustering algorithms in terms of accuracy and convergence speed. Lin *et al.* in the second article of this group argue that fast updated sequential pattern tree algorithms that normally update discovered sequential patterns in incremental mining require re-scanning of the original database. The authors propose an alternative algorithm that is based on the pre-large concept for maintaining discovered sequential patterns without rescanning the original database until the cumulative number of newly added customer sequences exceeds a safety bound. This results in improvements in the execution time for reconstructing the tree when old or new customer sequences are added into the original database to be reduced by using pre-large sequences. Their experiments reported in this article show the performance of the proposed algorithm for various minimum support thresholds and ratios of inserted sequences. Katrutsa *et al*. in the eight's article of this issue present a clustering algorithm that is based on the metric concentration location where the algorithm uses a reduced matrix of pairwise ranks distances. The key feature of the proposed algorithm is that it doesn't need the exhaustive matrix of pairwise distances and is primarily designed to solve the protein secondary structure recognition problem. The algorithm is compared with k-modes and tested on different metrics and data sets. Martinez *et al.* in the last group of these articles present a novel latent Dirichlet allocation (LDA) based probabilistic graphical approach for modeling and analyzing fluorescent spectroscopy excitation-emission Matrices (EEMs). The authors show that LDA-based model can increase classification performance, especially when paired with parallel factor analysis which may be regarded as perhaps the most popular and widely used tool for dealing with EEMs. Their experiments show that the proposed LDA-based algorithm is in some cases more robust to certain types of noise and data disturbances.

The last three articles of this issue are about various forms of raking and handling relational data. Du *et al.* in the tenth article of this issue argue the deficiencies of one of the most popular hyperlinking algorithms and to overcome the topic drifts, they propose a novel page ranking algorithm that combines the hyperlink with the triadic closure theory by considering the Vector Space Model (VSM) and the TrustRank algorithm. In their experiments, they use five classic HITS (Hyperlink-Induced Topic Search) based algorithms to compare with their proposed page ranking algorithm where they demonstrate that their proposed algorithm outperforms the four classic algorithms and the HITS algorithm, which is most commonly used. Makarehchi in the eleventh article of this issue argue that a major task in text categorization systems is dimensionality reduction, where among common methods, feature ranking-based feature selection, is scalable, simple, and inexpensive. However, selecting the most appropriate feature ranking method for a given data set without conducting experiments would be a challenge. The author presents a framework which is called feature meta-ranking, and is intended to identify the best feature ranking measure among a set of candidate solutions for a particular text classification problem. The proposed method is evaluated by applying it to six data sets. Seven feature ranking measures are employed and evaluated in this article. And finally in the last article of this issue, Tang *et al.* argue that a wide range of methods have been proposed for detecting different types of outliers. However, the interpretability of

outliers, that is, explaining in what ways and to what extent an object is an outlier, remains a critical issue. The authors propose an approach that is intended for improving this deficiency, develop a notion of multidimensional contextual outliers and propose a framework for contextual outlier detection. Here a contextual outlier is a small group of objects that share strong similarity with a significantly larger reference group of objects on some attributes, but deviate dramatically on some other attributes. An extensive set of experiments are presented in the paper to evaluate their proposed outlier detection approach.

In conclusion, with this issue of the IDA journal, which is Volume 19(5), we are pleased to see a consistent increase in the submission of high quality manuscripts to our journal. In addition to our six regular issues that now contain 11–12 articles, we expect to publish one special issue each year, which is normally related to a scientific event for which organizers have submitted an interesting proposal. 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