Intelligent Data Analysis 22 (2018) 935–937
DOI 10.3233/IDA-180893
IOS Press

Editorial

Dear Colleague:
Welcome to volume 22(5) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, the fifth issue of our twenty second year of publication, 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 four articles of this issue are about various aspects of advanced data preprocessing. Zou *et al.* in the first article of this group discuss the topic of oversampling in imbalanced data, discuss a well-known method, called SMOTE, that was introduced several years ago and propose two improved techniques based on SMOTE through sparse representation theory. Their extension results in replacing the k-nearest neighbors of the SMOTE with sparse representation, and using a sparse dictionary to create a synthetic sample directly. Their experiments performed on a number of UCI datasets show that both proposed methods can achieve better performance on TP-Rate, F-Measure, G-Mean and AUC values. The second article of this group by Guo et al is also about imbalanced data. The authors propose a simple but effective ensemble learning method based on feature projection and under-sampling that learns an ensemble through two steps: (1) under-sampling several subset from majority class and (2) constructing new training sets by projecting the original training set to different spaces defined by the matrixes. Their experimental results show that, compared with other state-of-the art methods this approach performs significantly better on measures of g-mean, f-measure, AUC, recall and accuracy. Hsu and Wu in the third article of this group emphasize that visualization is a useful technique in data analysis, where high-dimensional data is not visible and dimensionality reduction techniques are usually used to reduce the data to a lower dimension for visualization. Since most real-world datasets are of mixed-type containing both numeric and categorical attributes, the authors argue that a traditional approach could neither handle it directly nor output appropriate results. To address this problem, the authors propose a procedure for visualized analysis of mixed-type data via dimensionality reduction. Their proposed approach identifies significant features and visualizing patterns from the projection map chosen according to quality measures. Their experiments on real-world datasets were conducted to demonstrate feasibility of the proposed method. In the last article of this group Huertas *et al.* also argue the need for proper data reduction and introduce a HeatMap Based Feature Ranker, an algorithm an approach that is suitable for estimating feature importance purely based on its interaction with other variables. In this approach, a compression mechanism reduces evaluation space up to 66% without compromising efficacy. Their experiments show that their proposal is very competitive against popular algorithms, producing stable results across different types of data. The authors also show how noise reduction through feature selection aids data visualization using emergent self-organizing maps.

The next five articles are about various forms of unsupervised and supervised learning. Zhang *et al.* in the first article of this group explain that many contemporary data sources in a variety of domains can naturally be represented as fully-dynamic streaming graphs and how to design an efficient online streaming clustering algorithm on such graphs could be a challenge. The authors argue that existing clustering approaches are inappropriate for this task and propose a more suitable streaming clustering model consisting of a streaming reservoir and a cluster manager. The idea is based on an evolution-aware

bounded-size clustering algorithm to handle the edge additions/deletions. Their experimental results show that the proposed algorithm outperforms current online algorithms and is capable to keep track of the evolution of graphs. Bruneau and Otjacques in the next article of this group explain that model selection in spectral clustering is really estimating the ground truth number of clusters and propose a novel probabilistic framework where the spectral clustering pipeline relies on a latent representation over which a mixture model with K components is eventually fitted. The authors also propose an adapted Gaussian likelihood expression, and use it to derive a probabilistic model selection criterion for spectral clustering. The performance of the proposed method is evaluated on real and synthetic data sets, and compared to previous approaches in model selection for spectral clustering from the literature. Sun *et al.* in the seventh article of this issue discuss the topic of multi-label classification, explain the problem of excessive training time that restricts the availability of non-linear kernel and present a fast multi-label SVM training. The authors present extensive experiments on four large-scale benchmark data sets where their results show that the proposed algorithms can effectively reduce training time and their classification performance is similar to that of the traditional multi-label SVM algorithm. Serpen and Aghaei in the next article of this issue present the design and performance evaluation of a host-based misuse intrusion detection system for the Linux operating system. Their proposed system employs a feature extraction technique based on principal component analysis, which is called Eigentraces, of operating system call trace data, and k-nearest neighbor algorithm for classification. Classification of system call trace data that is in the form of feature vectors which are formulated through the Eigentraces procedure is accomplished using the k-nearest-neighbor algorithm. The authors evaluate two variants of the misuse intrusion detection system through a simulation study on the ADFA-LD dataset. In both cases their proposed design demonstrated very high performance and in overall, the misuse intrusion detection system was able to detect the attacks and predict the type of the attacks. Zarei *et al.* in the ninth article of this issue discuss the topic of finite population Bayesian bootstrapping in high-dimensional classification and argue that when the sample size is equal to or less than the number of covariates, traditional logistic regression is plugged with degenerates and wild behavior. To solve the problem, the authors use finite population Bayesian bootstrapping for resampling, such that the new sample size becomes greater than the number of covariates. They combine original samples and the mean of simulated data. The authors compare the proposed algorithm with the regularized logistic models and other popular classification algorithms using both simulated and real data.

The last group of articles in this issue are about novel methods in Intelligent Data Analysis. Pouralizadeh *et al.* in the first article of this group argue that implied volatility modeling is the future direction in price actuation and so has a crucial role in option pricing. The authors propose a machine learning polynomial approach due to the smile shaped behavior of implied volatility and investigate it with a regularization penalty term to fit the Out-The-Money volatility data and compare their results with the prominent counterpart method. Finally, their results illustrate that the new proposed algorithm yields an implied volatility smile which is free from static arbitrage for Out-The-Money European call options most of the time. Liu *et al.* in the next article of this group argue that the analysis of communities and their evolutionary behaviors in dynamic social networks is a challenging topic. The authors propose an approach for social community evolution that is based on gravitational relationship and community structure in a dynamic social network. The proposed method that concentrates on both community structure and micro-blog content can be used to reveal and track the process of community evolution. Their experimental results show that the proposed method performs well on detecting communities as well as tracking communities evolution in dynamic social networks. And finally Farzi *et al.* in the last article of this issue argue that word reordering is one of the main problems in machine translation and present a

neural reordering model based on phrasal dependency tree for statistical machine translation. Their proposed model that combines the power of the lexical reordering and syntactic pre-ordering is integrated into a standard phrase-based statistical machine translation system to translate input sentences. The authors evaluate their approach on syntactically divergent language-pairs, English-Persian and English-German using WMT07 benchmark. Their results illustrate the superiority of the proposed method in terms of precision and recall values with respect to the hierarchical, lexicalized and distortion reordering models.

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