

## Editorial

---

Dear Colleague:

Welcome to volume 24(1) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal is the first issue for our 24<sup>th</sup> year of publication. It contains eleven 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 are about data preprocessing in IDA. The first article by Kang and Oh is about balanced training/test set sampling for classification. They argue that random sampling does not guarantee that test accuracy reflects the performance of a developed classification model. The authors show the problems of random sampling and propose balanced sampling as an alternative. They also propose a measure for evaluating sampling methods and perform empirical experiments using benchmark datasets to verify that their sampling algorithm produces proper training and test sets. Their results confirm that their proposed method produces better training and test sets than random and several non-random sampling methods. The second article by Lv *et al.* is about Deep Neural Network (DNN) Models that are based on Dimensionality Reduction Operation (DRO). They argue that in order to avoid missing representative features, one should select a lot of features when using machine learning algorithms in stock trading while these high dimensional features can lead to redundancy of information and reduce the efficiency, and accuracy of learning algorithms. The authors select large-scale stock datasets from American and Chinese markets for their research project. For each stock dataset, they apply four most widely used DROs to deal with the original features, and then use the new features as input for the most popular DNN algorithms to generate trading signals. Their experiments show that only one method can significantly improve the performance of their modeling and the DRO does not significantly improve trading performance and the speed of generating trading signals. Tarawneh *et al.* in the last article of this group discuss feature selection and dimensionality reduction in deep learning for content-based image retrieval (CBIR). The authors compare the performance of CBIR systems using different deep features with state-of-the-art low-level features using different dictionaries and coefficient learning techniques. Their experimental results demonstrate a high level of mean average precisions when using deep features.

The second group of articles in this issue are about advanced learning in IDA. Yang *et al.* in the first article of this group discuss label distribution in ensemble learning. They argue that an ensemble of multiple member's forecasts can make better predictions while multiple individual models are generated to represent an uncertain system. The authors conduct data analysis based on the expertise of human forecasters and introduce a machine learning method for ensemble forecasting. Their experimental testing is performed on both artificial data and the data set for ensemble forecasting, where their results show that, compared with a baseline method and two state-of-the-art machine learning methods, their approach shows significantly better performance on measures of RMSE and average continuous ranked probability score. Cheng *et al.* in the fifth article of this issue discuss distance measuring for mixed data based on deep relevance learning. The authors propose an End-to-End Distance Measuring method for mixed data and argue that existing methods confuse the attributes space by mapping the discrete attribute values to

new continuous values, without considering the relevance. Their experimental results on a number of real-world datasets demonstrate that E2DM outperforms state-of-the-art methods. The next article by Sun *et al.* is about a latent-label denoising method for relation extraction with self-directed confidence learning. The authors propose a self-directed confidence learning based latent label denoising method for distantly supervised relation extraction. Their approach is based on a self-directed algorithm that combines the semantic information of model prediction and distant supervision to predict the confidence score of latent labels. Their experimental results show that their method can correct noisy labels with high accuracy and outperforms the state-of-the-art relation extraction systems. Wang *et al.* in the next article discuss community detection in dynamic networks using constraint non-negative matrix factorization. The authors argue that improving the community detection performance by combining network topology information in a short period is a challenging problem where previous efforts on utilizing such properties are insufficient. The authors introduce the geometric structure of a network to represent the temporal smoothness in a short time and propose a novel Dynamic Graph Regularized Symmetric NMF method to detect the community in dynamic networks. Their extensive experiments on multiple synthetic networks and two real world datasets demonstrate that the proposed method outperforms the state of the art algorithms on detecting dynamic community.

And finally the third group of articles are about novel methods in IDA. The first article of this group by Bauder and Khoshgoftar is about fraud prediction in medical claims data. The authors argue that although big data sources often contain a plethora of useful information but, in some cases, finding what is actually useful can be quite problematic and in binary classification problems, such as fraud detection, a major concern therein is one of class imbalance. They assess the impacts of class rarity in big data, and apply data sampling to mitigate some of the performance degradation caused by rarity. Their experiments show that rarity significantly decreases model performance, but data sampling, specifically random undersampling, can help significantly with rare class detection in identifying Medicare claims fraud cases. The next article of this group by Yu *et al.* is about extreme learning machine. The authors present a novel method which can objectively identify the subjective perception of tonic pain from EEG data. The method is first proposed for accurately extracting features of tonic pain from captured EEG data. Then, a single hidden layer feedforward network is used as a classifier for pain identification, and with the aid of extreme learning machine algorithm, the classifier is trained. Their experiments show the superiority of their classifier where they compare their results with a well-known support vector machine method. Guo *et al.* in the tenth article of this group discuss Storyline Extraction from News Articles with Dynamic Dependency. The authors argue that existing unsupervised approaches to storyline generation are typically based on probabilistic graphical models where they assume that the storyline distribution at the current epoch depends on the weighted combination of storyline distributions. The authors propose a new Dynamic Dependency Storyline Extraction Model in which the dependencies among events in different epochs but belonging to the same storyline are dynamically updated to track the time-varying distributions of storylines over time. The proposed model has been evaluated on three news corpora and the experimental results show that it outperforms the state-of-the-art approaches and is able to capture the dependency on historical contextual information dynamically. The last article of this group by Michi *et al.* is about optimization techniques for real-time bidding in advertising campaigns. The authors argue that while it is relatively easy to start an online advertising campaign, obtaining a high Key Performance Indicator (KPI) can be challenging. The authors propose an algorithm for advertisers to add an optimization layer which maximizes the chosen KPI by optimally configuring the entire process. Synthetic market data is used to evaluate the proposed approach against other state-of-the-art approaches adapted from similar problems.

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. For our volume 24(1), we are working on another special issue for which we can provide more details in the next two issues. 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