Editorial

Dear Colleague:

Welcome to volume 25(2) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal is the second issue for our 25th year of publication. This issue of contains thirteen articles representing a wide range of topics related to the theoretical and applied research in the field of Intelligent Data Analysis.

The first group of articles are about data preprocessing and advanced learning methods in IDA. Huang *et al.* in the first article of this group present an anomaly detection algorithm that is based on principal component analysis. The authors argue that traditional PCA-based detection algorithms commonly obtain a high false alarm for the outliers. To address this issue, the authors introduce the median and the median absolute deviation to rescale each outlier score that mapped onto the corresponding principal direction. And then, the true outlier scores of instances can be obtained as the sum of weighted squares of the rescaled scores. Their experimental results show that the proposed method has a good performance for effectiveness, efficiency, and robustness. The second article of this issue by Ye *et al.* is about a quadratic hyper-surface kernel-free least squares support vector regression. The idea here is to find a quadratic function as the regression function, which is obtained by solving a quadratic programming problem with the equality constraints. Essentially, the new model just needs to solve a system of linear equations to achieve the optimal solution instead of solving a quadratic programming problem. Compared with the standard support vector regression, their approach is much efficient due to kernel free and solving a set of linear equations. Their numerical results illustrate that the proposed approach has better performance than other existing regression approaches in terms of regression criterion and CPU time. Liu *et al.* in the next article argue that approximate multi pattern matching is an important issue and present a suffix array for multi pattern matching with variable length wildcards. The authors present two algorithms that are based on an efficient data structure for exact string matching as well as approximate pattern matching and multi pattern search applications. Their experimental results demonstrate that these algorithms, in most cases, are more time efficient than the state of the art comparison algorithms. Lit *et al.* in the next article of this group present a parametric approximation algorithm that is suitable for spatial group keyword queries. Their main motivation is the problem that efficient algorithms for solving these queries can only provide approximate solutions, and most of these algorithms achieve a fixed approximation ratio. To obtain a self-adjusting algorithm, the authors propose an approximation algorithm for achieving a parametric approximation ratio. The efficiency and scalability of the proposed algorithm are presented using experiments on benchmark datasets. Silva *et al.* in fifth article of this issue present a model to estimate the Self-Organizing maps grid dimension for prototype generation. The authors argue that despite its high accuracy, KNN has some weaknesses, such as the time taken by the classification process, which is a disadvantage in many problems, particularly in those that involve a large dataset. The authors propose a model that allows the best grid dimension of self-organizing maps and the ideal number of prototypes to be estimated using the number of dataset examples as a parameter. The main advantage of their proposed method is tested using eighteen public datasets which shows a better relationship between a reduced number of prototypes and accuracy, providing a sufficient number

that does not degrade KNN classification performance. Du *et al.* in the next article present an approach called deep multiple non-negative matrix factorization for multi-view clustering which is based on an auto-encoder. The approach consists of multiple encoder and decoder components with deep structures. Each pair of components are used to hierarchically factorize the input data from a view for capturing the hierarchical information, and all encoder and decoder components are integrated into an abstract level to learn a common low-dimensional representation for combining the heterogeneous information across multi-view data. Their experiments on six benchmark datasets demonstrate the superior performance of their proposed approach for multi-view clustering compared to other baseline algorithms. Similarly, Luo *et al.* in the next article of this group present a multi task prediction model that is based on ConvLSTM and encoder decoder. The authors argue that the energy load data in the micro-energy network is a time series with sequential and nonlinear characteristics. This model is applied to forecast the multi-energy load data of the micro-energy network in a certain area of China. Their test results prove that their model is convergent, and the evaluation index value of the model is better than competing methods. The last article of this group by Wan *et al.* is a similarity-based sales forecasting which is also an improved ConvLSTM and a complementary method. With the idea of collaborative filtering, the authors propose a similarity-based sales forecasting method which includes three modules. The approach is an attention-based ConvLSTM model which optimizes its loss function with the convex function information entropy. Their experimental results show that the proposed method can simultaneously adapt to the sales forecasting of mature products and new products, with a high accuracy of sales forecasting.

The second group of articles in this issue are about novel methods and enabling techniques in IDA. In the first article of this group, Huang *et al.* argue inferring user interest over large-scale microblogs have attracted much attention in recent years. However, the emergence of the massive amount of data, dynamic change of information and persistence of microblogs pose challenges to interest inference. The authors propose a novel user-networked interest topic extraction approach in the form of subgraph stream for microbloggers' interest inference. Their experimental evaluation on a large dataset from Sina Weibo, demonstrates that the proposed approach outperforms the state-of-the-art baselines in terms of precision, mean reciprocal rank as well as runtime from the effectiveness and efficiency perspectives. The second article of this group by Ge *et al.* is about a deep spatial-temporal fusion network for fine-grained air pollutant concentration prediction. The authors propose a general approach to predict air pollutant concentration, which consists of a data completion component, a similar region selection component, and a deep spatial-temporal fusion network. Their extensive experiments on a real-world dataset demonstrate that their model achieves the highest performance compared with state-of-the-art models for air quality prediction. In the next article, Zhang *et al.* introduce a novel method for time series similarity search using binary code representation and hamming distance. Their approach is able to represent original data compactly and can handle shifted time series and work with time series of different lengths. Moreover, it can be performed with reasonably low complexity due to the efficiency of calculating the hamming distance. Their experiments show that the proposed approach achieves better or has a comparative performance than the state-of-the-art methods in terms of accuracy and is much faster than most existing algorithms. Mao *et al.* in the next article present an automatic image detection applied to multi type surface defects on wind turbine blades which is based on a cascade deep learning network. This is a visual inspection model that can automatically and precisely classify and locate the surface defects, through the utilization of a deep learning framework based on the cascade R-CNN. The adaptability and generalization of their proposed model are validated by several types of defects and where their experiments demonstrate its superiority over the existing approaches with excellent results. And finally, Ning *et al.* in the last article of this issue present an adaptive node embedding framework for multiplex networks which is based on

cross-layer sampling strategies of nodes for multiplex networks. The approach is based on a fixed-length queue to record previously visited layers, which can balance the edge distribution over different layers in sampled node sequence processes. Their experiments on real-world networks in diverse fields show that their framework outperforms the state-of-the-art methods in application tasks such as cross-domain link prediction and mutual community detection.

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. Over the last few years, our submission rate has exceeded 600 manuscripts per year, with an acceptance rate of around 12–15%. 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