

## Editorial

---

Dear Colleague:

Welcome to volume 27(2) of the Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal is the second issue for our 27<sup>th</sup> year of publication. It contains fourteen 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 in this issue are about state of the art data pre-processing methods in IDA. Li and Lu in the first article of this issue introduce a self-training algorithm that is based on globally adaptive multi-local noise filter. The idea is based on density peaks for empowering self-training to label unlabeled samples. Their experimental results conducted on eighteen UCI data sets demonstrate that their proposed approach is not sensitive to the value of the neighbor parameter  $k$ , and it is capable of adaptively finding the appropriate number of neighbors of each class. The second article of this group by Wang et al. is about a novel feature selection method that is based on feature relevance, redundancy and interaction in neighborhood rough set. The authors introduce a new method of information measurement called neighborhood symmetric uncertainty, to measure what proportion of data a feature contains regarding category label. The results on the nine datasets and five representative feature selection algorithms are presented by the authors. The next article by Zhou et al. is also about feature selection in which the authors introduce a grouping feature selection method that is based on feature interaction. The authors also introduce a new evaluation function measuring feature interaction and a grouping strategy that is based on approximate Markov blanket. Their experimental results on fifteen public data sets show that their proposed approach outperforms others in terms of classification accuracy and Macro-F1. Cheng et al. in the fourth article of this issue present an approach for improving multi-label learning by modelling local label and feature correlations. Multi-label learning deals with the problem that each instance is associated with multiple labels simultaneously. Many methods have been proposed for modeling label correlations in a global way to improve the performance of multi-label learning. Their experimental results on twelve real-world multi-label data sets demonstrate the effectiveness of the proposed method. In the next article of this issue, Wang et al. introduce a Relief-PGS algorithm for feature selection and data classification. The idea is to overcome the shortcomings of SVM (Support Vector Machine) where the penalty factor and kernel function of SVM and the extracted feature of Relief algorithm are encoded as the particles of particle swarm optimization-genetic algorithm and optimized by iteratively searching for optimal subset of features. Numerical experimental results indicate that the classification accuracy and efficiency of the proposed approach are superior to those other algorithms including traditional SVM. In the last article of this group Strnad et al. introduce a synthetic dataset generator for anomaly detection that is suitable for university environment. The generator is Opensource and is able to scale particular class of data, time-wise and also perform injection of the data with cyber-attackers' behaviour patterns. Different types of real attack behaviour patterns in the university environment have been selected and are used to demonstrate attackers' behaviour in synthetically created system logs. The mentioned features allow other researchers to benchmark their anomaly detection algorithms with complex data.

The second group of articles are about supervised and unsupervised learning methods in IDA. In the first article of this group, Santos and Campos introduce a supervised clustering algorithm with attributed

networks. The authors argue that an increasing area of study for economists and sociologists is the varying organizational structures between business networks. They introduce a new method of supervised clustering with attributed networks. One of the innovative aspects of their proposed approach is the use of a supervised clustering algorithm to attributed networks that can be accomplished through a combination of weights between the matrix of distances of nodes and their attributes when defining the clusters. The proposed methodologies are applied to an inter-organizational network for which they present their results. Chen et al. in the eighth article of this issue introduce a domain density peak clustering algorithm based on natural neighbour. Some of the advantages of this approach are: it is sensitive to the cutoff distance; the neighborhood information of the data is not considered when calculating the local density; and during allocation, one assignment error may cause more errors. A series of experiments are conducted that demonstrate higher accuracy and robustness of the proposed approach. Wang et al. in the ninth article of this issue present an intrusion detection algorithm that is based on transfer extreme learning machine. The authors argue that at present, most of the existing intrusion detection methods are based on traditional machine learning algorithms. These methods need enough available intrusion detection training samples, where training and test data should meet the assumption of independent and identically distributed patterns. Their experiments are carried out on several public data sets where the experimental results show that the algorithm can improve the detection accuracy, especially for unknown and small samples. Li et al. in the next article of this group present an approach for integration of deep neural network with logic rules for credit scoring which is an important topic in financial activities and bankruptcy prediction. The proposed framework calculates the rule satisfaction distance for each instance using a probabilistic soft logic formula. In addition the logic rules are integrated into the posterior distribution of the DNN output to form a logic output. Finally, a novel discrepancy loss, which measures the difference between the real label and the logic output, is used to incorporate logic rules into the parameters of the neural network. Their extensive experiments that were conducted on two datasets, show that compared to the standard model, the four evaluation metrics are increased by a reasonable amount.

The third group of articles are about enabling techniques and innovative case studies in IDA. Delahoz-Domínguez et al. in the first article of this group discuss application of machine learning in the ex-combatant demobilization process on the Colombian armed conflict. The authors explore the potential of supervised machine learning models to support the decision-making process in demobilizing ex-combatants in the peace process in Colombia. The proposed approach makes a significant contribution by training and evaluating four machine learning models, using a database composed of many individuals and a set of selected variables. From their results, it was possible to conclude that the XGBoost algorithm is the most suitable for predicting the future status of an ex-combatant. The twelfth article of this issue by Xi and Xu is about design of a dynamic Gaussian deep belief network and its application to stock market. The authors propose a model of Dynamic Gaussian Deep Belief Network. In their experiments the forecast for the stocks of large industrial corporations is compared with DBN and LSTM. In the next article of this issue Flyckt et al. explain rifle shooting factors through multi-sensor body tracking. The authors argue that many studies have correlated body posture and balance to shooting performance in rifle shooting tasks, but have mostly focused on single aspects of postural control. This study has focused on finding relevant rifle shooting factors by examining the entire body over sequences of time. The dataset and pre-processing pipeline, as well as the techniques for generating explainable predictions presented in this study have laid the groundwork for future research in the sports shooting domain. And finally, the last article of this issue by Chen et al. is about multi-granularity user interest modeling and interest drift detection in which the authors argue that the traditional service model based on the search engine can no longer meet the increasing demand for personalized service. The authors propose a hierarchical classification tree, named

HAT-tree, to maintain the history of the user's preferences at multi-topic and multi-time granularity. Their proposed algorithm can find the user's long term and short-term preferences, detect the user's explicit and implicit preference drift, and highlight the importance of the user's more recent preferences. The article includes experiments carried out on multiple data sets, and the experimental results show that the proposed method is more accurate than other similar algorithms of user preference drift detection.

In conclusion for the second issue of our volume 27, I would like remind you that as the founding Editor-in-Chief of the IDA journal, I am gradually wrapping up my duties and I have transferred the responsibility to my colleague, Dr. Jose Maria Pena (from Oxford, UK), whom I have known since 1997. Please join me in welcoming Dr. Pena to the position of the Editor-in-Chief of the IDA Journal. We are also glad to announce that our impact factor has increased by over 50% since last year (from 0.860 to 1.321). 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*      *Dr. J.M. Pena*  
*Founding Editor*   *Editor-in-Chief*