Intelligent Data Analysis 23 (2019) 1187–1189
DOI 10.3233/IDA-190006
IOS Press

Editorial

Dear Colleague:

Welcome to volume 23(6) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal, is the sixth and the last issue for our 23rd year of publication. It 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 two articles are about learning from imbalanced data in IDA. Yun and Sun in the first article of this issue focus on the problem of sequential prediction for imbalanced data streams in which the authors propose a novel hybrid algorithm for handling this class of problem. The proposed method is able to perceive the changeable data distribution and perform the adaptation by itself, thus building a reliable model with a low fitting deviation. Their experiments conducted on different kinds of UCI datasets demonstrate that the proposed algorithm not only has better generalisation performance but also provides higher numerical stability. Hou *et al.* in the second article of this issue argue that ensemble learning is an excellent method for imbalance classification. They emphasize that existing ensemble methods often ignore noise in the dataset, and propose a density-based undersampling algorithm (DBU) where they integrate it with AdaBoost (DBUBoost) to improve the classification performance. The authors introduce a similarity coefficient to distinguish the examples from each category. To demonstrate the effectiveness of their proposed method, they compare DBUBoost with four ensemble methods and three anti-noise methods. Their experimental results have shown that DBUBoost performs better than other state-of-the-art methods.

The second group of articles in this issue are about advanced learning methods in IDA. Li *et al.* in the first article of this group argue that uncertain data that are accompanied with probability make frequent itemset mining more challenging. Their article is about the problem of mining probabilistic maximal frequent itemsets in which they re-define the concept of probabilistic maximal frequent itemset to be consistent with the traditional definition and provide a better view on how to devise pruning strategies. Their theoretical analysis and experimental studies demonstrate that their proposed algorithms have high accuracy, expend less computational time use less memory, and significantly outperform the state-of-the-art algorithm. Saifan and Al-Smadi discuss source code-based defect prediction using deep learning and transfer learning. They argue that discovering errors and fixing defective software modules early in the project lifecycle (e.g. in the testing phase) can save resources and enhance software quality. The authors utilize multiple datasets to initialize a Deep Belief Network model and transfer the obtained knowledge to train their discovered model using a source project in a cross-project combination. Their evaluation of 13 open Java projects from several repositories shows that their proposed model achieves improvements based on several performance measures. Kalintha *et al.* in the fifth article of this issue propose a novel distance metric learning called evolutionary distance metric learning (EDML) to improve clustering quality that simultaneously evaluates inter- and intra-clusters. The authors empirically demonstrate the drawback of EDML in non-linearly separable input space and illustrate the benefit of kernel function to the extension K-EDML method by showing its superior result benefits to other clustering algorithms in the semi-supervised clustering on various real-world datasets. Bellodi *et al.* in the next article of this group argue that although recent advances of subgraph mining enable us to find subgraphs that are statistically significantly associated with the class variable from graph databases, it is challenging to interpret the resulting subgraphs due to their massive number and their propositional representation. The authors represent graphs by probabilistic logic programming and solve the problem of summarizing significant subgraphs by structure learning of probabilistic logic programs. They further empirically demonstrate that their approach can effectively summarize significant subgraphs with keeping high accuracy. Qu *et al.* in the seventh article of this issue emphasize that in cyber anomaly detection, if the detected target is significantly different from the predefined normal network data pattern, it is considered an outlier. However, the degree of deviation from the normal model is often difficult to determine, making it challenging to effectively identify attack categories that are similar to normal network data and have small sample sizes. To address this problem, the authors propose a novel anomaly detection method called a comparison network (C-Net), which has a double-branch structure for a neural network. They performed experiments using a water storage dataset. Their models detection rate of the Complex Malicious Response Injection (CMRI) attack category reached 95.5%, while the cyber anomaly detection algorithms based on machine learning could not detect any attacks. As for the KDDCUP99 data, their model achieved a very high detection.

And finally the third group of articles are about enabling techniques in IDA. The first article of this group by Guo *et al.* is about analysing drug-target interaction cluster analysis based on improving the density peaks clustering algorithm. The authors argue that since drug-target data have neither class labels nor the cluster number information, they are not suitable for clustering algorithms that require predefined parameters determined by comparing clustering results with real class labels. The authors consider density peaks clustering (DPC) that can determine the number of clusters without requiring class labels and propose an improved local density method based on a cut-off distance sequence that overcomes the limitations of DPC and can be successful in analysing drug-target data. Their drug-target data clustering results of the improved algorithm are more reasonable than the results of the fast K-medoids and hierarchical clustering algorithms. Ruiz *et al.* in the ninth article of this group discuss credit scoring for microfinance using behavioral data in emerging markets. They argue that despite the enormous number of inhabitants, these financial markets still lack a proper finance infrastructure where the main difficulties felt by customers is access to loans. This limitation arises from the fact that most customers usually lack a verifiable credit history. The authors propose credit scoring modeling based on non-traditional-data, acquired from smartphones, for loan classification processes they use Logistic Regression and Support Vector Machine models which are the top linear models in traditional banking. Their models surpassed the performance of the manual loan application selection process, improving the approval rate and decreasing the overdue rate and compared to the baseline, the loans approved by meeting the criteria of the SVM model presented a decreased overdue rate. Aghababaei and Makarehchi in the next article discuss an interpolative self-training approach for link prediction where they propose learning social networks from incomplete relationship data. The authors address link prediction as a semi-supervised learning problem where the task is to predict a larger part of networks using available knowledge of smaller parts. They propose an interpolative self-training technique that leverages node information to generate a set of examples in learning phase along with their connections as their associated labels. This approach generates data by interpolation of documents assigned to a pair of nodes. To evaluate their proposed method, the authors perform a set of experiments on co-authorship networks of 18 different domains. Their results imply the feasibility of achieving significantly high performance for most of the networks using the proposed self-training approach. Wang *et al.* in the eleventh article of this issue argue that he traditional trajectory privacy protection algorithm approaches the task as a single-layer problem.

Taking a perspective in harmony with an approach more characteristic of human thinking, in which complex problems are solved hierarchically, the authors propose a two-level hierarchical granularity model for this problem. The authors theoretically prove that the proposed algorithm outperforms the traditional algorithm in terms of data distortion and anonymity cost and verify its efficacy experimentally. And finally Tang *et al*. in the last article of this issue explain collaborative filtering (CF), one of the most famous methods for building recommendation systems, which recommends relevant items to users or predicting ratings of users' unknown items. The authors introduce an online-and-offline collaborative filtering with a multi-method model to improve the traditional CF method, called Online SGD with Offline Knowledge (OSGDO for short). Their method proves to be good at online training when new observations arrive. And the results of their experiments show that the dynamic training process they propose is more efficient than rebuilding the model on all the data.

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