DOI 10.3233/IDA-190004
IOS Press

Editorial

Dear Colleague:
Welcome to volume 23(4) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal is the fourth issue for our 23$^{\text{rd}}$ 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 three articles in this issue are about advanced data preprocessing in IDA. Li and Gao in the first article argue that data gathered from real world application often contains label noise, which is harmful to the quality of data where most data mining processes suffer a deterioration when they are applied to noisy data. The authors propose to improve data quality by correcting mislabeled data and employing a procedure to estimate the level of noise in the data and combining this noise estimation with the correction process. Their experimental results using real-world data sets from UCI machine learning repository show that their approach successfully improves data quality in many cases and outperforms several existing correction methods. Du *et al.* in the next article of this issue tackle the problem of outlier detection and how to promptly detect the local outlier of a large-scale mixed attribute data in the big data era. The comprehensive experiments performed on large-scale benchmark datasets show that the performance of their approach, which is based on deleting the massive clear non-noise and extracting cluster-based pre-noise set, is superior to the existing methods. This is when it is compared to the state-of-the-art techniques. Park *et al.* in the third article of this issue argue that many domain adaptation methods for NLP have been developed on the basis of numerical representations of texts instead of textual inputs. The authors develop a distributed representation learning method of words and documents for domain adaptation. The authors propose a new method based on negative sampling through learning document embeddings by assuming that a noise distribution is dependent on a domain. Among several experiments, the authors verified that their proposed method outperformed other representation methods in terms of indiscriminability of the distributions of the document embeddings through experiments such as visualizing them and calculating a proxy A-distance measure.

The second group of articles in this issue are about unsupervised learning in IDA. Hsu *et al.* in the first article of this group discuss that to facilitate handling of categorical values SOM algorithm has been extended in order to incorporate data structure distance hierarchies. This was done by taking into account the semantics embedded in categorical values via distance hierarchies. This has been based on supervised learning which demands presence of a class attribute in the dataset where in real-world applications, class attribute may not be available. The authors present several methods of unsupervised learning of distance hierarchies where neither class attribute nor domain experts required in measuring similarity degrees between categorical values. Their experiments verify feasibility and performance of the proposed unsupervised-learning approach. Ming *et al.* in the next article of this group argue that k-means clustering is the most popular clustering technique, where Lloyd's algorithm is the most popular algorithm of the k-means due to its simplicity, geometric intuition and effectiveness. However, even in Lloyd's algorithm, one needs to compute the Euclidean distances between all data points and all cluster centres in each iteration which prevents the algorithm from being scalable to large datasets. The authors propose two scalable k-means algorithms, which are data-parallel techniques in order to scale beyond computational and memory limits of a single machine. Their Extensive experiments conducted on four large-scale datasets show that their proposed algorithms have good convergence performance and achieve almost ideal speedup.

The third group of articles are about various forms of classifications in IDA. Chen *et al.* in the sixth article of this issue argue that in the context of expensive and time-consuming acquisition of reliably labelled data, how to utilize the unlabelled instances that can potentially improve the classification accuracy becomes an attractive problem with significant importance in practice. The authors propose a self-learning framework, that firstly pre-learns a classification model using the labelled data, then makes the prediction of unlabelled instances in the form of soft class labels. Their experiments demonstrate that the semi-supervised models can produce better generalization accuracy than the supervised counterparts. Kasemtaweechok, and Suwannik in the next article of this group explain as the training data becomes larger in any application, the KNN algorithm suffers from drawbacks such as large storage requirements, slow classification speed, and high sensitivity to noise. The authors propose a novel prototype selection technique based on geometric median and compare it with seven state-of-the-art prototype selection methods. Their experiments show that their proposed method runs faster than the baseline model and the classification accuracy and kappa value of the proposed method are comparable to those of all considered state-of-the-art prototype selection methods. Wei *et al.* in the last article of this group explain that network representation learning aims at learning a low-dimensional vector for each node in a network, in which most existing approaches only use topology information of each node and ignore its attributes information. The authors propose an Improved Attributed Node Random Walks (IANRW) framework, which constructs the neighborhood of an attributed node and then leverages the skip-gram model to perform node embedding. Their experiments on six datasets show that IANRW outperforms many state-of-the-art embedding models and can improve various attributed networks mining tasks.

And finally, the last four articles in this issue are about enabling methods in IDA. Garcia-Rudolph *et al.* in the ninth article of this issue argue that in patients with Traumatic Brain Injury predictive techniques have traditionally been applied for cognitive rehabilitation gross outcome prognosis (e.g. cognitive improvement or not), without considering treatment configuration variables. The authors propose to enrich predictive models with variables that therapists can act upon. The authors report superior performance to previous state-of-the-art models with similar datasets and suggest use cases including the obtained predictive models that contribute to treatments personalization and efficiency. Ma *et al.* in the next article of this group emphasize the importance of personalised decision making, such as personalised medicine and online recommendations, and argue that most existing methods assume a known cause (e.g. a new drug) and focus on identifying from data the contexts of heterogeneous effects of the cause. The authors argue that there is no approach to efficiently detecting directly from observational data context specific causal relationships, and propose a Tree based Context Causal rule discovery method, for exploration of context specific causal relationships from data. Their experiments with both synthetic and real world data sets show that their proposed approach can effectively discover context specific causal rules from the data. Tang *et al.* in the next article argue that measuring semantic textual similarity (STS) lies at the core of many applications in natural language processing where most models have considered semantic information or syntactic information, but seldom a unified model to make full use of these two kinds of information. The authors propose a semantic-embedded dependency tree model based on word2vec and glove, which can be treated as a syntactic semantic representation. The authors have applied the method to textual similarity tasks and evaluated its performance through two widely used benchmarks: the Pearson correlation coefficient and the Spearman correlation coefficient. Their experimental results show that SEDT/E-SEDT can effectively improve the accuracies of sentence similarity judgments. And finally, Li *et al*. in the last article of this issue discuss that for the Bayesian network (BN) structure learning, the key problem is to determine the relationship between the BN nodes. In this article the authors propose the scheme of group decision making based on the intuitionistic fuzzy set for the relationship determination between the BN nodes. The authors apply their proposed scheme to establish the model for the thickening process of gold hydrometallurgy where they conclude that the expert who owns bigger membership degree and less hesitancy degree plays the most important role in the process of decision making.

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