Intelligent Data Analysis 23 (2019) 971–973
DOI 10.3233/IDA-190005
IOS Press

Editorial

Dear Colleague:

Welcome to volume 23(5) of Intelligent Data Analysis (IDA) Journal.

This issue of the IDA journal is the fifth 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 four articles are about various forms of data preprocessing in IDA. Zhou *et al.* in the first article of this issue discuss outlier detection in data analysis. The authors introduce a novel outlier detection algorithm which integrates the local density with the global distance seamlessly. In the proposed method, an integrated outlier factor is used to measure the detecting accuracy. Their experimental study on both synthetic and real-life datasets shows that the proposed method is more effective than some typical outlier detection methods, such as Relative Density-based Outlier Score, INFLuenced Outlierness, Local Outlier Factor and Local Distance-based Outlier detection Factor. Li *et al.* in the second article of this issue discuss the topic of curse of dimensionality and the complicated correlation among dimensions where dimension reduction methods often are used to alleviate these problems. The authors propose an outlier detection method based on Variational Autoencoder, which combines low-dimensional representation and reconstruction error to detect outliers. Their experiments conducted on six real-world datasets show that their proposed method performs better than or is at least comparable to state of the art methods. Chakroun *et al.* in the next article of this issue explain some guidelines for enhancing data locality of IDA. They argue that for ML algorithms to produce results in a reasonable amount of time, they need to be implemented efficiently. The authors analyze one of the means to increase the performances of machine learning algorithms which is exploiting data locality. They start by motivating why and how a more efficient implementation can be achieved by exploiting reuse in the memory hierarchy of modern instruction set and further document the possibilities of such reuse in some selected machine learning algorithms. Praczyk in the last article of this group compares a number of state-of-the-art anomaly detection methods on ship trajectories obtained by an Automatic Identification System (AIS) in the Baltic sea. Because most methods need fixed length trajectory representations, this paper also gives some solutions for reducing variable length trajectories to a fixed size.

The second group of articles in this issue are about unsupervised learning in IDA. Peng *et al.* in the first article of this group propose a novel segmentation and clustering-based identification algorithm to effectively recognize fixations, saccades and smooth pursuits in eye tracking. In their proposed algorithm, the authors employ the velocity feature in the recorded eye data to identify the saccade segments, and then the standard deviation of the dispersion is used to divide the remaining data into segments. The authors evaluate the proposed algorithm with the eye tracking dataset sampled from 11 participants where their experimental results show that the proposed mechanism can achieve high accuracy and good recall. Oladipupo and Olugbara in the next article of this group argue that the widely used data analytics based clustering algorithms are highly data dependent, making it pertinent to find the most effective algorithm for knowledge mining in any dataset associated with student engagement. The authors evaluate

the performances of five famous clustering algorithms in which the k-means algorithm is benchmarked with 22 distance functions based on the Silhouette index, Dunn's index and partition entropy internal validity metrics. The overall ranking of the clustering algorithms was based on cluster potentiality using the median deviation statistics. The results of their evaluation show the well-known k-means algorithm to have the highest cluster potentiality, demonstrating its effectiveness for the task of knowledge mining in a student engagement dataset. Large *et al.* in the seventh article of this issue discuss the topic of time-series classifications with dictionary-based classifiers where a sliding window is used across each series, discretising the window to form a word, forming a histogram of word counts over the dictionary, then constructing a classifier on the histograms. The authors explain that a significant difference could exist in accuracy between these seemingly similar algorithms and investigate this phenomenon by deconstructing the classifiers and measuring the relative importance of the four key components between different methods. Their experiments show their proposed improvements would significantly improve the classification accuracy. Liu *et al.*, in the last article of this group, also discuss text classification where they emphasize that classification of sentences is very challenging, since sentences contain limited contextual information. The authors propose an Attention-Gated Convolutional Neural Network for sentence classification, which generates attention weights from the feature's context windows of different sizes by using specialized convolution encoders. Their experimental results show that their model can achieve higher accuracy than standard CNN models, and gain competitive results over the baselines on four out of six tasks that they compared.

And finally the third group of articles are about advanced learning techniques in IDA. Diaz-Pacheco and Reyes-Garcia in the first article of this group explain that full model selection is a technique to improve the accuracy of machine learning algorithms through the search of the most appropriate combination on each dataset for feature selection, data preparation, applied learning algorithm and the adjustment of its hyper-parameters. The authors also emphasize that although this paradigm has been widely studied in large datasets, it has been poorly explored in high volume datasets. The authors propose the use of the full model selection paradigm to construct proxy models. Their obtained results, show a performance without significant differences in comparison with the complete search algorithm. Bahrami and Sajedi in the next article of this group argue that concept detection for a collection of images is an important topic where it is normally facing an imbalanced dataset with great challenges. To cope with this challenge, the authors propose an image concept detection system based on the Convolutional Neural Network method? Using a large image dataset, their experimental results demonstrate the effectiveness of the proposed framework for concept detection in imbalanced datasets. Rnjbar-Sahraei *et al.* in the eleventh article of this issue argue that to extract structured knowledge from unstructured text sources involves understanding the semantic relationships between entities. However, in domains with sparse data such as social networks which have limited occurrences of entities and relationship patterns, bootstrapping techniques and pattern detection methods are inefficient and inaccurate. The authors introduce a Relation Extraction approach based on Distant Supervision which extracts the named entities from text documents and assigns a fingerprint to each potential relationship among the named entities. The authors evaluate their proposed approach on a non-English historical archive consisting of unstructured notarial acts and structured civil registers where it achieves precision of 0.90. And finally, Wang *et al.* in the last article of this issue discuss the topic of location privacy preservation and argue that existing location-privacy-preserving methods primarily focus on solving the problem of location-privacy preservation in the global space. This not only increases the response time of the location service, it also degrades the data quality. The authors propose a k-anonymity algorithm based on locality-sensitive hashing which achieves higher efficiency and higher quality of service through a bottom-up grid-search method. The

results of their experiments conducted on the proposed algorithm indicate a smaller anonymous spatial region, higher data quality and lower time cost than methods with no subspace.

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. Our special issue for 2019 was recently published from some of the best papers of FSDM-2018 conference that was held in Bangkok, Thailand. We are also working on another special issue for 2020 to include the best papers from CIARP-2019 conference that will be held later this year in Havana, Cuba. 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