Intelligent Data Analysis 28 (2024) 1407–1410
DOI 10.3233/IDA-249006
IOS Press

Editorial

Dear Colleague:

Welcome to volume 28(6) of the Intelligent Data Analysis (IDA) Journal.

Dear reader, welcome to the sixth issue of IDA, which is in its 28th year. This issue contains a collection of papers related to different theoretical and applied topics in Intelligent Data Analysis. The first part of the issue presents theoretical contributions, while the second part includes several application papers in different areas.

We start the first part with two papers focused on unsupervised learning. Fei Zhang, et al. present an unsupervised divide-and-conquer method for identifying manipulated profiles in recommender systems, explicitly targeting standard and obfuscated behavior attacks. By categorizing profiles and clustering them in the extracted feature space without requiring prior knowledge or annotations, the method enhances precision and simplicity in detection. Experimental results on MovieLens-100K and Netflix demonstrate the model's superior accuracy and reduced running time compared to existing detection methods. This approach addresses the limitations of traditional attack detection techniques by providing a tailored solution for different types of shilling attacks. In the second paper on this topic byTran, et al., the authors introduce a POCS-based clustering algorithm that leverages the convergence properties of the Projection onto Convex Sets (POCS) method to achieve effective data clustering. The algorithm treats each data point as a convex set and performs simultaneous projections of cluster prototypes onto member data points, utilizing adaptive weights to minimize a predefined objective function. Extensive experiments demonstrate that the proposed algorithm outperforms traditional clustering methods such as K-means/K-means++ and Fuzzy C-Means (FCM) in both effectiveness and efficiency. The results affirm the practical applicability and validity of the POCS-based clustering approach for various clustering tasks.

The second group of theoretical contributions are centered on graph mining and analysis. First, the paper by X. Liu, et al. introduces KGNN_HCD, a novel community detection method for heterogeneous graphs that addresses the limitations of traditional models, which either overlook feature space information or require extensive domain knowledge for defining meta-paths. The proposed method constructs a K-nearest neighbor graph using a similarity matrix, generates a meta-path information matrix through a weighted convolutional transformation, and employs a graph convolutional network (GCN) for high-quality node representation. Extensive experiments on three datasets (ACM, DBLP, and IMDB) show that the KGNN_HCD method significantly improves community detection performance, outperforming 11 existing methods. The results indicate that KGNN_HCD is both effective and applicable to complex network classification and clustering tasks. The second graph-related paper is authored by Yuan et al.. In this paper, the authors introduce an online course evaluation model leveraging a graph auto-encoder to address the limitations of previous methods that overlook the correlations among large-scale online learning behavior data. By constructing K-Nearest Neighbor (KNN) graphs from collected course data and employing a variational graph auto-encoder (VGAE) to learn implicit features, the model enhances understanding of student behaviors. Experimental results on two datasets demonstrate significant improvements, with a more than tenfold increase in the Calinski-Harabasz index for clustering tasks and approximately 10% better performance in classification tasks than traditional methods. These findings highlight the model's effectiveness in analyzing complex educational data for online course evaluation.

We conclude the theoretical contributions with three contributions on feature selection and engineering. In Z. Liu, et al., the authors introduce a chaos-based binary dragonfly algorithm (CBDA) for feature selection. This algorithm incorporates three innovative improvements – chaotic mapping, evolutionary population dynamics, and a binarization strategy – to enhance the conventional dragonfly algorithm's exploration and exploitation capabilities. Experiments conducted on 24 datasets from the UCI repository demonstrate that the CBDA outperforms established algorithms in terms of fitness value, classification accuracy, CPU running time, and the number of selected features. The second paper on this topics is by Göcs & Johanyák. In their paper, they discuss the preprocessing and feature selection workflow for intrusion detection systems (IDSs), emphasizing the importance of selecting a minimal set of features to distinguish between malicious and benign network traffic effectively. Utilizing the CSE-CIC-IDS2018 dataset on AWS, six feature selection methods were applied, resulting in a final ranking of features based on their average scores. Various subsets of features were created according to different ranking thresholds and tested with five classification algorithms to determine the optimal feature set for five specific attack types. The evaluation considered four widely used metrics, demonstrating the workflow's effectiveness in enhancing IDS performance. We finish the feature-related group of papers with the paper by Noronha & Zarate, in which they address the challenges of characterizing longevity profiles from longitudinal data using the English Longitudinal Study of Ageing (ELSA-UK). The authors employ feature engineering techniques such as merging, factor analysis, and biclustering to select relevant features that effectively discriminate between long-lived and non-long-lived individuals to reduce high dimensionality. Two classification models – one based on a decision tree and the other on a random forest – are developed from the preprocessed data, demonstrating successful discrimination of longevity profiles. The results reveal that the economic situation and mobility correlations significantly impact longevity, suggesting that this methodology could be beneficial for analyzing other longitudinal studies.

The second part of this issue contains papers with a substantial applied contribution. We start with an application paper on temporal modeling for the semantic web. In this paper, Fu Zhang, et al. tackles an exciting topic in the temporal analysis of the semantic web. The paper introduces an efficient index for querying bitemporal RDF (Resource Description Framework), which accommodates more complex situations involving both valid time and transaction time, overcoming the limitations of current static RDF representations. By innovatively incorporating a redesigned skip list structure, the proposed index aims to cover a wide range of query patterns while minimizing the number of indexes needed. Additionally, it considers the performance of standard RDF queries when the time element is not defined. Experimental results using the Lehigh University Benchmark (LUBM) demonstrate that the proposed index is both scalable and effective for handling temporal RDF data. The following two papers focus on NLP and, in particular, sentiment and opinion analysis. First, Huang, et al. introduce BEMOJI, a pre-training model based on Bidirectional Encoder Representations from Transformers (BERT) that incorporates emojis for enhanced sentiment analysis in both Chinese and English. The model improves performance by predicting emoji descriptions from related texts during pre-training and using a fusion layer to combine text representations and emoji descriptions in sentiment predictions. Experimental results indicate that BEMOJI significantly outperforms traditional emoji-based and transformer-based methods, achieving high accuracy and macro metrics. Additionally, the robustness studies confirm that BEMOJI performs comparably to BERT on tasks without emojis, underlining its effectiveness and reliability. The second paper on this topic, by Feng, et al., explores the impact of extreme agents on opinion evolution in social networks, highlighting their significant influence compared to regular agents, which often undermines the development of opinion neutrality. By introducing a temporal dimension and the concept of time sunk cost, the research categorizes agents into four states based on their opinions, focusing on the effects of extreme state agents. The findings reveal that restricting the exchange of opinions among extreme agents can significantly reduce their numbers by 40% to 50%, and implementing these restrictions early on increases the likelihood of shifting network opinions towards neutrality. Overall, the results underscore the influential role of extreme agents in shaping the dynamics of opinion extremization in online platforms. We conclude the NLP-oriented papers with an interesting contribution on speech recognition where Fan, et al. introduce Sampleformer, an optimized version of the Convolution-augmented Transformer (Conformer) model, which has achieved state-of-the-art results in Automatic Speech Recognition (ASR). The authors identify inefficiencies in the original Conformer design and implement downsampling in the Conformer Encoder to enhance model efficiency and accuracy. A novel multi-group attention mechanism is also proposed, significantly reducing attention complexity. Experimental results on the AISHELL-1 dataset demonstrate that the 13.3 million-parameter Sampleformer achieves a 3.0%/2.6% reduction in character error rate and offers 30% faster inference and 27% quicker training compared to the baseline Conformer model.

The following two papers are focused on applied image analysis. In the first one, Jiménez et al. focus their work on reducing speckle noise in breast ultrasound images while preserving essential features using two GAN models: Conditional GAN (CGAN) and Wasserstein GAN (WGAN). The models were trained on public breast ultrasound databases, and their performance was evaluated through Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Results indicated that the CGAN model outperformed WGAN in despeckling performance, achieving a PSNR of 38.18 dB and an SSIM of 0.96. The findings suggest that the superior performance of CGAN can enhance future computer-aided detection and diagnosis systems for breast cancer. In the second paper on this subject, Yu, et al. present a Multi-local Feature and Attention fused network (MFA) designed for person re-identification (ReID) to address challenges such as local occlusion, scale misalignment, and attitude changes in pedestrian images. The proposed system utilizes a Channel Point Affinity Attention module to enhance local detail extraction, and it segments feature maps for focused attention on different pedestrian areas through a Global Local Aligned loss function. The MFA model achieved significant performance on datasets such as Market-1501 and DukeMTMC-reID. Additionally, the model operates efficiently at approximately 32 frames per second and demonstrates competitive performance compared to other ReID methods.

We conclude this issue with two more applied contributions, this time focused on sustainability and environmental data analysis. In the paper by Gupta, et al., the authors present their research on detecting crop diseases in tomatoes, soybeans, and mushrooms using real-time and publicly available datasets characterized exclusively by categorical attributes, presenting unique challenges. After encoding labels and addressing missing values through four preprocessing techniques, the study employs the SMOTE-N technique to handle class imbalance, followed by classification using bagging, boosting, and voting ensemble methods, optimized with the Ant Lion Optimizer for hyper-parameter tuning. The evaluation of twelve models reveals that the hybrid model II-SN-OXGB, which combines Random Forest for imputation and optimized Xtreme Gradient Boosting for classification, outperforms all others in classification accuracy across thirteen categorical datasets. The findings indicate that applying this model can significantly enhance crop disease detection, allowing farmers to implement timely interventions to mitigate yield losses and economic impacts. We conclude this part and the issue with Liu, et al.'s paper, which addresses the growing concern over sulfur dioxide (SO2) emissions in China, predominantly generated by fossil-fired power plants, and proposes a hybrid deep learning model for predicting these emissions during the flue gas desulphurization (FGD) process. The model integrates a temporal convolution neural network (TCNN) and a gated recurrent unit (GRU) with a mutual information technique to effectively select relevant variables and capture the complex dynamics of SO2 emissions. Utilizing data from an actual

1000 MW coal-fired power plant, the model demonstrates superior predictive performance compared to existing methods across various performance indicators. The findings suggest that this approach can effectively enhance the modeling of SO2 emissions in dynamic FGD processes.

With our best wishes,

Dr. A. Famili	***Dr. J.M. Peña***
Founder	***Editor-in-Chief***