

Editorial

Dear Colleague:

Welcome to volume 28(4) of the Intelligent Data Analysis (IDA) Journal.

Dear reader, welcome to this fourth issue of our 28th year of the IDA journal. In this issue, we have selected papers covering different theoretical and applied topics in the field of Intelligent Data Analysis.

In the first half of the issue, we cover mostly theoretical or methodological contributions in both algorithms and techniques, whilst for the second part we have selected several application papers in different areas.

We start the issue with Modarres' paper on outlier detection methods, which discusses the impact of outliers on distance matrices, showing how outliers can affect pairwise distances and inflate eigenvalues. It proposes a new outlier detection method based on eigenvalues and compares it with existing techniques across various distributions, applied to different real-world problems.

The next two papers present contributions in the area of unsupervised learning. Hosszú's paper addresses the topic of pattern evolution research within the context of unsupervised and dimensionality reduction methods and proposes a novel method for examining tested pattern systems to confirm their classification, which is based on a composite multivariate method that can be used in the evolutionary research of pattern systems. This method allows to check the feature selection made by the earlier successive elimination method and split the features into two groups: those created during the pattern systems' initial development phase from a shared ancestor, and those made later. In the second paper of this topic, Johnson, Giraud-Carrier, & Hatch, present a novel approach in time-series classifications in graphs. The article presents a new inductive bias for learning in dynamic event-based human systems by incorporating dynamic contrastive learning in pre-training, without relying on polynomial expansions or universal approximators. This approach, demonstrated on event-based graph time-series classification, addresses the challenge of deep learning in chaotic systems by assuming that the relationship between input features and output classes evolves over time, resulting in improved performance on real-world data.

The next three contributions address problems in image and computer vision. First, Zhang et al., introduce a new approach for the 3D human motion prediction problem based on Multiple Distilling-based spatial-temporal attention (MD-STA) networks to extract temporal and spatial features respectively and fuse them. The proposed architecture, uses three modules (Screening Self-attention (SSA) and Frames and Keypoint-Distilling (FKD), and Dim-reduction Fusion (DRF)). In the second paper, Tian and Cheng's contribution discusses the development of a model for unsupervised domain adaptation (UDA) in person re-identification. The new proposed method leverages multiple source domains, employs sample weighting to address sample imbalances, and uses adversarial learning to align domains, demonstrating effectiveness across multiple datasets. The third paper on image analysis, presented by Wang et al., introduces a lightweight method for 6D pose estimation of indoor objects' point clouds by mobile robots, addressing challenges such as domain gap and high computational costs. By employing an enhanced PointNet++ network structure and lightweight modules to create a codebook, the proposed model achieves impressive validation results on YCB-Video and LineMOD datasets. The study demonstrates that the model effectively estimates the position and pose of unknown object point clouds with reduced

computational and storage requirements, outperforming high-precision methods in terms of parameter efficiency and real-time performance.

The last of the sections on theoretical and methodological contributions covers the topics of text analysis and NLP. In the first paper of this block, Wang et al., address the importance of Named Entity Recognition (NER) in Natural Language Processing and highlights the effectiveness of utilizing a mix of character-word structure and dictionary information for Chinese NER. The proposed method, ELCA, addresses challenges such as long-distance entities and detection of multiple entities with the same character, achieving state-of-the-art results in Chinese Word Segmentation and NER by incorporating sentence-level position information and adaptive word convolution. Zhang et al., in the second paper of this topic, tackle an interesting computational linguistics problem. The paper details an innovative hierarchical feature decoupling model for generating SQL queries from natural language in text-to-SQL tasks. By separating features for subtasks within SELECT and WHERE clauses, the model improves performance significantly, outperforming state-of-the-art baseline methods on the WikiSQL benchmark dataset. We conclude the section in text analysis methods and the theoretical/methodological papers with He et al., who present their contribution in text classification using LSTM networks. The article introduces a BiLSTM-GCN hybrid neural network text classification model based on dependency parsing to improve news topic text classification accuracy by incorporating semantic information. By combining BiLSTM for feature extraction, dependency parsing for semantic relationships, GCN for global information, and a global average pooling layer to prevent overfitting, the proposed method achieved impressive results on well-know dataset problems, outperforming traditional text classification methods.

The second part of the issue include several papers in different application fields. The first one is on time-series analysis. Zhang et al., propose an Attention-Embedded Time-Aware Imputation Network (ATIN) with two sub-networks is proposed for the classification of anomalous data from production time series in the oilfield affects future analysis and forecasting. The method introduced in the paper uses a Time-Aware Imputation LSTM (TI-LSTM) network to model irregular time intervals and incomplete measurements while a second Attention-Embedding LSTM (ATEM) is designed to improve the effectiveness of anomaly detection. In the second application paper, Meng et al., present a new CNN model called MusicNeXt for music genre classification, focusing on capturing temporal information to distinguish fused music genres effectively. The proposed model enhances feature extraction by utilizing temporal information more effectively and introducing a genre-sensitive adjustment layer to increase distinctiveness between genres, outperforming baseline networks and other state-of-the-art methods in music genre classification tasks without generating category bias in the results.

The next two papers corresponds to a second subsection in application papers focus on finance or marketing applications. First, Malik et al.'s paper analyzes the helpfulness of online product reviews by integrating robust contextual word embeddings, topic, and language models. By employing various feature generation techniques and a wrapper-based feature selection method, the paper concludes that the ELMo model surpassed standard baselines and even outperformed the fine-tuned BERT model in predicting review helpfulness on Amazon datasets for Video games and Health & personal care. Additionally, the LDA model demonstrated comparable performance to BERT but outperformed other baseline models, showcasing the framework's ability to reveal crucial factors in product reviews and its potential for evaluation across different platforms. In the second paper on finance, Chen et al., delve into utilizing deep forests, a novel approach combining neural networks and ensemble learning, for credit fraud detection. Introducing the distributed dense rotation deep forest algorithm (DRDF-spark) based on RotBoost, the model addresses spatial correlation challenges in data, employs Spark for parallel construction to enhance processing speed, and includes a pre-aggregation mechanism for improved communication efficiency.

Experimental results demonstrate the superiority of DRDF-spark over deep forests and mainstream ensemble learning methods on fraud detection datasets, showcasing up to more than three times faster training speeds and the potential for further speedup with increased nodes.

We conclude this issue with the last application paper; Cao et al., present an interesting paper on cybersecurity. The paper introduces a robust Log parsing method based on Self-supervised Learning (LogSL) to extract templates from logs of varying formats without requiring labels. By using a Multi-token Prediction Model (MPM) combining different modules, LogSL achieves higher parsing accuracy compared to existing methods in analyzing system logs for anomaly detection and security tasks.

With our best wishes,

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