10th IEEE International Conference on Data Science & Advanced Analytics (DSAA)

IEEE | Co-sponsors ACM/ASA/CCF

DSAA 2023

October 9-12, 2023
Thessaloniki, Greece

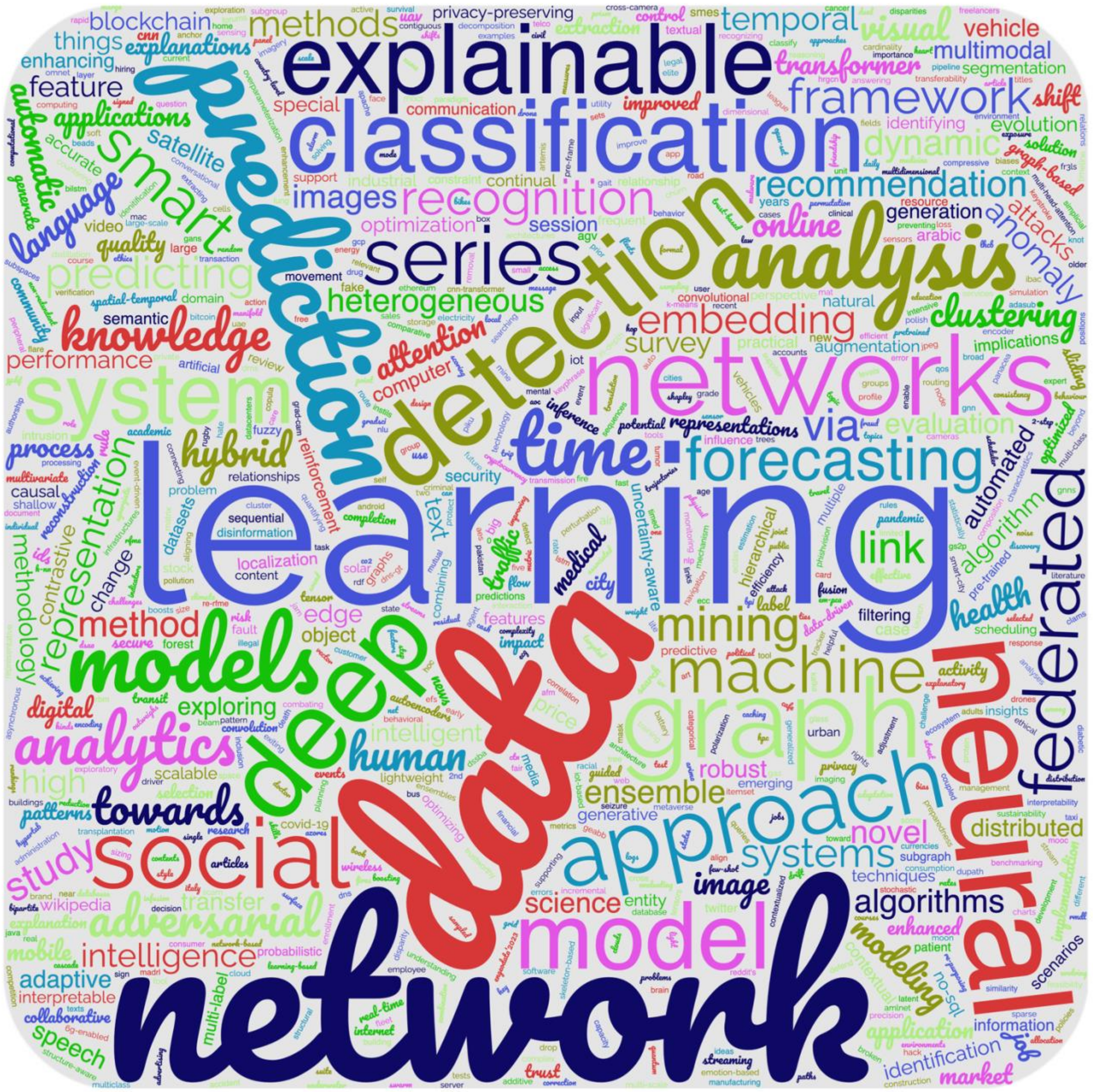
https://conferences.sigappfr.org/dsaa2023/

IEEE Advancing Technology for Humanity

IEEE Computational Intelligence Society

OpenCEMS Connected Environment & Distributed Data Management Solutions

Keywords’ cloud



Contents

DSAA'23 Committee Message

The 10th IEEE International Conference on Data Science & Advanced Analytics (DSAA-2023) was held in Thessaloniki, Greece from 9 to 12 of October in person, as before the pre-covid period. DSAA-2023 was co-organized by the Aristotle University of Thessaloniki and the Open University of Cyprus, whereas it was sponsored by the IEEE. Previous DSAA events took place at Shenzhen, China (2022), Porto, Portugal (2021), Sydney, Australia (2020), Washington DC, USA (2019), Turin, Italy (2018), Tokyo, Japan (2017), Montreal, Canada (2016), Paris, France (2015), Shanghai, China (2014).

The 10th IEEE International Conference on Data Science & Advanced Analytics (DSAA-2023) features its strong interdisciplinary synergy between statistics, computing and information/ intelligence sciences, and cross-domain interactions between academia and business for data science and analytics. DSAA sets up a high standard for its organizing committee, keynote speeches, submissions to main conference and special sessions, and a competitive rate for paper acceptance. DSAA has been widely recognized as a dedicated flagship annual meeting in data science and analytics such as by the Google Metrics and China Computer Foundation. DSAA-2023 provides a premier forum that brings together researchers, industry and government practitioners, as well as developers and users of big data solutions for the exchange of the latest theoretical developments in Data Science and the best practice for a wide range of applications. DSAA-2023 invited submissions of papers describing innovative research on all aspects of data science and advanced analytics as well as application-oriented papers that make significant, original, and reproducible contributions to improving the practice of data science and analytics in real-world scenarios.

The DSAA structure consists of several separate tracks with separate PC committees: Research track, Applications track, Industrial track, Journal track, Special sessions, and a Data Science Competition. In particular, the list of Special sessions is the following:

- PSTDA: Private, Secure, and Trust Data Analytics
- AISC: AI and Data Science for Cybersecurity
- PRAXAI: Practical Applications of Explainable Artificial Intelligence Methods
- DISA: Computational methods for emerging problems in disinformation analysis
- CIVIL: Computational Imaging, Vision, Linguistics and Language
- SmartCities: Smart City Data Analytics: Applications and Implications
- GraDSI: Graph Data Science and Applications
- DSSBA: Data Science for Social and Behavioral Analytics
- LfTD: Learning from Temporal Data

This year, we received 260 submissions from 30 countries (as shown in Figure 1) out of which only 118 full papers were accepted for oral presentation at the conference program and inclusion in the IEEE proceedings. In particular, the acceptance ratio for each of the main tracks was: 25% for the Research Track, 24% for the Applications Track, and 24% for the Industrial Track.

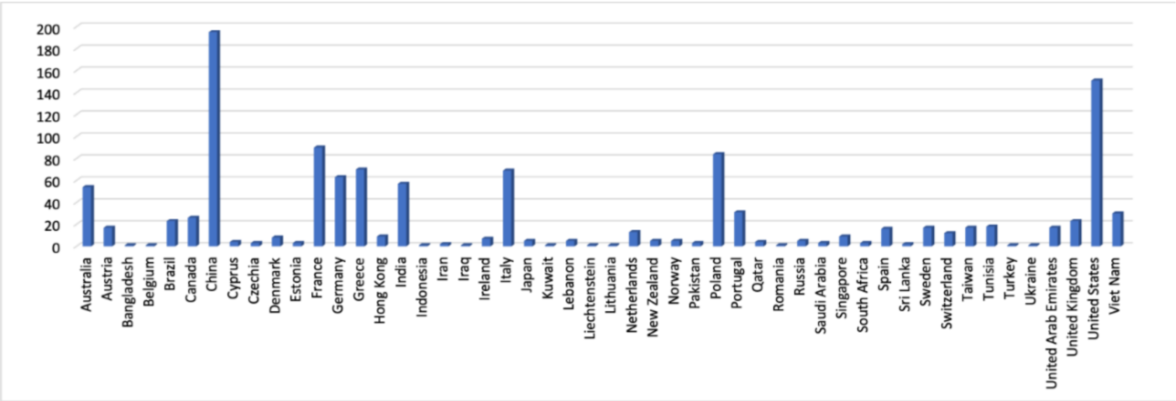


Figure 1. Countries of Submitted Papers.

Each paper was reviewed by at least 3 reviewers. As shown in Figure 2, the members of the Program Committee belong to not least than 28 countries.

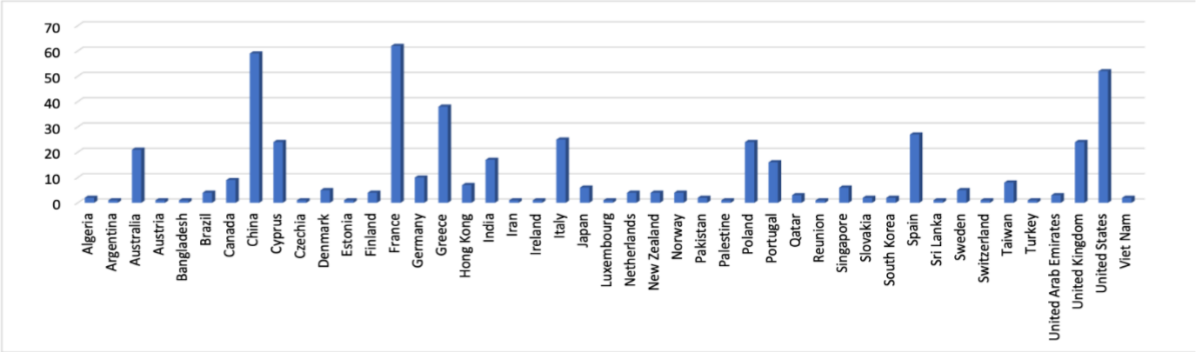


Figure 2. Countries of PC members of DSAA'23.

DSAA-2023 program includes four well-known keynote speakers:

- Sihem Amer-Yahia (University of Grenoble Alpes), who delivered a talk on "Towards AI-Powered Data-Informed Educational Platforms",
- Angela Bonifati (Lyon 1 University), who delivered a talk on "Towards Quality-driven and AI-assisted Data Science",
- Yannis Ioannidis (University of Athens), who delivered a talk on "User-Defined Functions in Relational Databases: Challenges and Promising Solutions based on YeSQL",
- Gerhard Weikum (Max Planck Institute for Informatics in Saarbruecken), who delivered a talk on "What Computers Know, and What They Should Know".

On the other hand, DSAA-2023 program included a tutorial given on:

- "Interpretability Methods for Graph Neural Networks", delivered by Arijit Khan and Ehsan Bonabi Mobaraki (Aalborg University).

We would like to thank our keynotes and our tutorialists for their contribution to the success and sustainability of DSAA.

We would like to thank all authors for submitting their papers to DSAA-2023 and we hope they will submit their research papers again in future DSAA events. On the other hand, we express our gratitude to all the Program Committee members who provided high-quality reviews. We want to acknowledge the ease of use and flexibility of the EasyChair system to manage papers. Finally, we would like to thank our sponsor, IEEE, as well as the constant support of the local organizers.

For conference attendants, we hope they enjoyed the technical program, informal meetings, and interaction with colleagues from all over the world.

DSAA 2023 Chairs

Yannis Manolopoulos, Open University of Cyprus, Cyprus
Zhi-Hua Zhou, Nanjing University, China
Guoliang Li, Tsinghua University, China
Timos Sellis, Archimedes/Athena RC, Greece
Minos Garofallakis, Technical University of Crete, Greece
Takashi Washio, Osaka University, Japan
Peter Triantafyllou, Warwick University, UK
Athena Vakali, Aristotle University of Thessaloniki, Greece
Bin Yang, Aalborg University, Denmark
Feida Zhu, Singapore Management University, Singapore

Venue

The conference was held at the **Grand Hotel Palace** in Thessaloniki, featuring several keynote speakers and authors who presented their fascinating research studies and contributions.

Grand Hotel Palace
Monastiriou 305
Thessaloniki, 54628 Greece

Contact: Phone: +30 2310 549 000   Mail: info@grandhotelpalace.gr

About the Grand Hotel Palace

Grand Hotel Palace is a five star hotel in Thessaloniki. Grand Hotel Palace is located at a central point at the entrance of the city and has a total capacity of 258 rooms and suites in modern or classic style distributed on the six floors of the neoclassical building of unique architecture. Grand Hotel Palace is the biggest Conference Hotel in Thessaloniki with 13 elegant and multi-purpose conference halls & meeting rooms. The mission of the people of Grand Hotel Palace is to offer high aesthetic hospitality in a luxurious and comfortable environment.

Grand Hotel Palace is the largest 5-star conference hotel in Thessaloniki; the city of history and culture and the evergreen economic centre of the Balkan region. Grand Hotel Palace combines neoclassical elegance with comfort, timeless luxury with functionality, and finesse with high-quality services.

Right in the heart of the philosophy of the Grand Hotel Palace lies respect for the environment, the local community, and the city's culture, as well as for our people, employees, partners, and guests. The mission of the Grand Hotel Palace is to provide premium-quality hotel services in the field of conference and leisure tourism, within an environment of high aesthetics. Our objective is to meet the visitor's quality expectations at all levels of accommodation and hospitality.

The neoclassical building of the Grand Hotel Palace rises impressively at a central point in the west entrance, just three and a half kilometres from the centre of Thessaloniki. The emblematic building of the hotel that was erected in 2004 reflects timeless neoclassical aesthetic value.

At Grand Hotel Palace, every detail has been carefully designed to meet all requirements, in terms of comfort and ambience. The hotel is an expression of elegance that reflects a graceful blend of architecture, consisting of 249 rooms and 9 suites in modern or classical style.

A conference space of 2500 m² with 13 fully renovated multi-functional conference rooms, infused with natural lighting and advanced technology can suit every occasion from international conferences and meetings to social events.

'Seros Restaurant' and the elegant 'Marco Polo' lounge bar offer delightful choices of high quality, as well as 'The Garden', the new specially designed area in the courtyard, while the Health Club is available for exercise and wellness.

For more information: https://www.grandhotelpalace.gr/

Sponsors & Organizers

IEEE Advancing Technology for Humanity

IEEE Computational Intelligence Society

Connected Environment & Distributed Energy Data Management Solutions

General Information

Registration Desk Opening Hours

Monday, October 9, 2023:	From 8:00 AM to 5:00 PM
Tuesday, October 10, 2023:	From 8:00 AM to 5:00 PM
Wednesday, October 11, 2023:	From 8:00 AM to 5:00 PM

Instructions for Session Chairs

When you chair a session

- Introduce yourself and the session papers.
- Ask the authors to show up.
- You might need to remind them about the timing of each paper type (between 15’ and 20’).
- Remind attendees to raise their hands for interactions with the presenter.
- Following each presentation, watch for attendees’ raised hands and allow them to interact with the presenter one by one.

Instructions for Speakers

All speakers are requested to come to the session room during the coffee break at least 1 hour in advance of their presentations to verify if the data will function properly on the equipment provided.

Please name the file as: “Presentation no. (or session name)-presenter name.ppt”.

To avoid virus infection, kindly scan your data with an updated anti-virus software be-forehand.

A Personal Computer will be available for presentations in each room.

Tickets

Delegates will receive their tickets (Gala Diner, Welcome Reception, Excursion, etc.) at registration desk. Tickets are to be displayed when needed. If you have misplaced your ticket or have not received tickets for the function you wish to attend, please visit the staff at the registration desk. Tickets are available for purchase subject to availability.

Urgent Messages

Urgent messages for delegates can be directed to the registration desk. Messages will be held at the registration desk for collection and the recipient will be notified via a notice board.

Name Tags and Luggage Lockers

Please wear your name tag at all time during the conference, including lunch and conference dinner. You may be asked to present your name tag.

Mobile Phones, Pager & Laptop Sound

As a courtesy to presenters and colleagues, please ensure that your mobile phones, pagers and laptop speakers are switched off during the conference sessions.

Internet Access

A wireless internet access is provided during the conference days.

About Thessaloniki

Thessaloniki, also known as Thessalonica, Saloniki, Salonika, or Salonica, is the second-largest city in Greece, with slightly over one million inhabitants in its metropolitan area, and the capital of the geographic region of Macedonia, the administrative region of Central Macedonia and the Decentralized Administration of Macedonia and Thrace. It is also known in Greek as “the co-capital”, a reference to its historical status as the “co-reigning” city of the Byzantine Empire alongside Constantinople.

Thessaloniki is located on the Thermaic Gulf, at the northwest corner of the Aegean Sea. It is bounded on the west by the delta of the Axios river. The municipality of Thessaloniki, the historical center, had a population of 319,045 in 2021, while the Thessaloniki metropolitan area had 1,006,112 inhabitants and the Thessaloniki (regional unit) had 1,092,919. It is Greece's second major economic, industrial, commercial and political center, and a major transportation hub for Greece and southeastern Europe, notably through the Port of Thessaloniki. The city is renowned for its festivals, events and vibrant cultural life in general, and is considered to be the country's cultural capital. Events such as the Thessaloniki International Fair and the Thessaloniki International Film Festival are held annually, while the city also hosts the largest bi-annual meeting of the Greek diaspora. Thessaloniki was the 2014 European Youth Capital. The city's main university, Aristotle University, is the largest in the Balkans.

The city was founded in 315 BC by Cassander of Macedon, who named it after his wife Thessalonike, daughter of Philip II of Macedon and sister of Alexander the Great. An important metropolis by the Roman period, Thessaloniki was the second largest and wealthiest city of the Byzantine Empire. It was conquered by the Ottomans in 1430 and remained an important seaport and multi-ethnic metropolis during the nearly five centuries of Turkish rule, and from the 16th to the 20th century was the only Jewish-majority city in Europe. It passed from the Ottoman Empire to the Kingdom of Greece on 8 November 1912. Thessaloniki exhibits Byzantine architecture, including numerous Paleochristian and Byzantine monuments, a World Heritage Site, as well as several Roman, Ottoman and Sephardic Jewish structures.

Thessaloniki is a popular tourist destination in Greece. In 2013, National Geographic Magazine included Thessaloniki in its top tourist destinations worldwide, while in 2014 Financial Times FDI magazine (Foreign Direct Investments) declared Thessaloniki as the best mid-sized European city of the future for human capital and lifestyle.

About Vergina

Vergina is a small town in Northern Greece, part of Veria municipality in Imathia, Central Macedonia. Vergina was established in 1922 in the aftermath of the population exchanges after the Treaty of Lausanne and was a separate municipality until 2011, when it was merged with Veria under the Kallikratis Plan.

Vergina is best known as the site of ancient Aegae, the first capital of Macedon. In 336 BC Philip II was assassinated in Aegae's theatre and his son, Alexander the Great, was proclaimed king. In 1977, the burial sites of several kings of Macedon were uncovered, including the tomb of Philip II, which had not been disturbed or looted, unlike so many of the other tombs there. The ancient town was also the site of an extensive royal palace. The archaeological museum of Vergina was built to house all the artifacts found at the site and is one of the most important museums in Greece.

Aegae has been awarded UNESCO World Heritage Site status as “an exceptional testimony to a significant development in European civilization, at the transition from classical city-state to the imperial structure of the Hellenistic and Roman periods”.

Social Events

Welcome Reception
Date: Monday, October 9, 2023
Time: 7-8 pm
Location: Grand Hotel Palace (Grace B Room)

Conference Banquet
Date: Tuesday, October 10, 2023
Time: 8-10 pm
Location: Makedonia Palace Hotel
Departure: 7:30 pm from Grand Hotel Palace

Excursion
Date: Thursday, October 12, 2023
Time: 2-7 pm
Trip to Vergina and visit of the archeological site.
Departure: 2 pm from Grand Hotel Palace

Full Conference Program

Monday 9th of October 2023

Monday 9:00 am – 9:30 am

Opening Ceremony
(*Room: Grace C*)

Monday 9:30 am – 10:30 am

Keynote Talk: Towards Quality-driven and AI-assisted Data Science
(*Room: Grace C*)
By: Angela Bonifati
Abstract: Page 29

Monday 10:30 am – 12:00 pm

Session 1: Advanced Analytics and Knowledge Discovery Methods (Research I)
(*Room: Grace C*)
(Abstracts: Page 35)

Designing Concept Drift Detection Ensembles: A Survey (PDF)

Martin Trat and Jivka Ovtcharova

Sliding Window Sampling over Data Stream — A Solution Based on Devil's Staircases (PDF)

Dominik Bojko, Jacek Cichoń and Mirosław Kutyłowski

slidSHAPs – sliding Shapley Values for correlation-based change detection in time series (PDF)

Chiara Balestra, Bin Li and Emmanuel Müller

Measurement of Illegal Android Gambling App Ecosystem From Joint Promotion Perspective (PDF)

Yadi Han, Shanshan Wang, Yiwen Li, Xueyang Cao, Limei Huang and Zhenxiang Chen

Monday 10:30 am – 12:00 pm

Session 2: Business and Industry (Applications I)
(*Room: Grace A*)
(Abstracts: Page 36)

Electricity Price Forecasting Based on Order Books: A Differentiable Optimization Approach (PDF)

Léonard Tschora, Tias Guns, Erwan Pierre, Marc Plantevit and Celine Robardet

Contextual Advertising Strategy Generation via Attention and Interaction Guidance (PDF)

Issam Benamara and Emmanuel Viennet

HRGCN: Heterogeneous Graph-level Anomaly Detection with Hierarchical Relation-augmented Graph Neural Networks (PDF)

Jiaxi Li, Guansong Pang, Ling Chen and Mohammad-Reza Namazi-Rad

FIW-GNN: A Heterogeneous Graph-based Learning Model for Credit Card Fraud Detection (PDF)

Kuan Yan, Junbin Gao and Dmytro Matsypura

Monday 10:30 am – 12:00 pm

Session 3: Private, Secure, and Trust Data Analytics (PSTDA I)
(*Room: Grace D*)
(Abstracts: Page 37)

Stochastic Perturbation Averaging Boosts Transferability of Adversarial Examples (PDF)

Rongbo Yang, Qianmu Li and Shunmei Meng

Novel Few-shot Learning Based Fuzzy Feature Detection Algorithms (PDF)

Yun Luo, Liangfu Lu, Xudong Cui, Yan Du, Yingying Bi, Limin Zhu and Christy Jie Liang

A Contextualized Transformer-Based Method for Cyberbullying Detection (PDF)

Nabi Rezvani, Amin Beheshti and Xuyun Zhang

Privacy-Preserving Learning via Data and Knowledge Distillation (PDF)

Fahim Faisal, Carson K. Leung, Noman Mohammed and Yang Wang

Monday 12:00 pm – 12:20 pm

Coffee Break
(*Room: Grace B*)

Monday 12:20 pm – 1:50 pm

Session 4: Advanced Classification Methods (Research II)
(*Room: Grace C*)
(Abstracts: Page 39)

Evaluating Explanation Methods of Multivariate Time Series Classification through Causal Lenses (PDF)

Etienne Vareille, Adel Abbas, Michele Linardi and Vassilis Chrsitopides

Interpretable Time Series representation for Classification Purposes (PDF)

Etienne Le Naour, Ghislain Agoua, Nicolas Baskiotis and Vincent Guigue

LSFuseNet: Dual-Fusion of Landsat-8 and Sentinel-2 Multispectral Time Series for Permutation Invariant Applications (PDF)

Arshveer Kaur, Poonam Goyal and Navneet Goyal

A Novel Method for Temporal Graph Classification Based on Transitive Reduction (PDF)

Carolina Jeronimo, Zenilton Patrocínio Jr., Simon Malinowski, Guillaume Gravier and Silvio Guimaraes

Monday 12:20 pm – 1:50 pm

Session 5: Business and Education (Applications II)
(*Room: Grace A*)
(Abstracts: Page 40)

Enhancing the Performance of Automated Grade Prediction in MOOC using Graph Representation Learning (PDF)

Soheila Farokhi, Aswani Yaramala, Jiangtao Huang, Muhammad Fawad Akbar Khan, Xiaojun Qi and Hamid Karimi

Supplier Qualification Document Recognition through Open-set Recognition (PDF)

Giuseppe Rizzo and Angelo Impedovo

Identifying Survival-Changing Sequential Patterns for Employee Attrition Analysis (PDF)

Youssef Oubelmouh, Frédéric Fargon, Cyril de Runz, Arnaud Soulet and Cyril Veillon

Towards Deep Learning Models for Automatic Computer Program Grading (PDF)

Peter Nagy and Heidar Davoudi

Monday 12:20 pm – 1:50 pm

Session 6: Private, Secure, and Trust Data Analytics (PSTDA II)
(*Room: Grace D*)
(Abstracts: Page 42)

Privacy-aware Adaptive Collaborative Learning Approach for Distributed Edge Networks (PDF)

Saeed Alqubaisi, Deepak Puthal, Joy Dutta and Ernesto Damiani

Multi-Granularity Entity Recognition Based Sentence Ranking for Multi-Document Summarization (PDF)

Guowei Zhang, Xuyun Zhang, Zhiyong Wang and Amin Beheshti

Temporal Differential Privacy for Human Activity Recognition (PDF)

Debaditya Roy and Sarunas Girdzijauskas

Graph Disentangled Collaborative Filtering Based on Multi-order Similarity Constraint (PDF)

Yaoze Liu, Junwei Du, Haojie Li and Guanfeng Liu

Monday 1:50 pm – 3:00 pm

Lunch Break

Monday 3:00 pm – 4:30 pm

Session 7: Time Series and Forecasting (Research III)
(*Room: Grace C*)
(Abstracts: Page 43)

Combining Forecasts using Meta-Learning: A Comparative Study for Complex Seasonality (PDF)

Grzegorz Dudek

Deep Spectral Copula Mechanisms Modeling Coupled and Volatile Multivariate Time Series (PDF)

Yang Yang, Zhilin Zhao and Longbing Cao

Spatial-Temporal Residual Multi-Graph Convolution Network for Traffic Forecasting (PDF)

Ruoxuan Zhu, Yi Qian, Hui Zheng, Xing Wang, Junlan Feng, Lin Zhu and Chao Deng

AMLNet: Adversarial Mutual Learning Neural Network for Non-AutoRegressive Multi-Horizon Time Series Forecasting (PDF)

Yang Lin

Monday 3:00 pm – 4:30 pm

Session 8: Private, Secure, and Trust Data Analytics (PSTDA III)

(*Room: Grace A*)

(Abstracts: Page 44)

Defending the Graph Reconstruction Attacks for Simplicial Neural Networks (PDF)

Huixin Zhan, Liyuan Gao, Kun Zhang, Zhong Chen and Victor Sheng

Underwater Localization Based on Robust Privacy-preserving and Intelligent Correction of Sound Velocity (PDF)

Jingxiang Xu, Ying Guo, Ziqi Wang, Fei Li and Ke Geng

A Multimodal Adversarial Database: Towards a Comprehensive Assessment of Adversarial Attacks and Defenses on Medical Images (PDF)

Junyao Hu, Yimin He, Weiyu Zhang, Shuchao Pang, Ruhao Ma and Anan Du

Enhancing Federated Learning by One-Shot Transferring of Intermediate Features from Clients (PDF)

Deng Youxingzhu, Zhou Yipeng, Liu Gang, Hui Wang and Shui Yu

Monday 3:00 pm – 4:30 pm

Session 9: AI and Data Science for Cybersecurity (AISC)

(*Room: Grace D*)

(Abstracts: Page 46)

CRIMEO: Criminal Behavioral Patterns Mining and Extraction from Video Contents (PDF)

Raed Abdallah, Hassan Harb, Yehia Taher, Salima Benbernou and Rafiqul Haque

Cross-layer Federated Heterogeneous Ensemble Learning for Lightweight IoT Intrusion Detection System (PDF)

Suzan Hajj, Joseph Azar, Jacques Bou Abdo, Jacques Demerjian, Abdallah Makhoul and Dominique Ginhac

A Data-driven Approach for Risk Exposure Analysis in Enterprise Security (PDF)

Albert Calvo, Santiago Escuder, Josep Escrig, Marta Arias, Nil Ortiz and Jordi Guijarro

Understanding the Country-Level Security of Free Content Websites and their Hosting Infrastructure (PDF)

Mohammed Alqadhi, Ali Alkinoon, Saeed Salem and David Mohaisen

ECC: Enhancing Smart Grid Communication with Ethereum Blockchain, Asymmetric Cryptography, and Cloud Services (PDF)

Raphaelle Akhras, Wassim El-Hajj, Hazem Hajj, Khaled Shaban and Rabih Jabr

Tuesday 10th of October 2023

Tuesday 9:00 am – 10:30 am

Session 10: Knowledge Graphs and Graph Learning (Research IV)
(*Room: Grace C*)
(Abstracts: Page 47)

Knowledge Graph-based Embedding for Connecting Scholars in Academic Social Networks (PDF)

Prasad Calyam, Xiyao Cheng, Yuanxun Zhang, Harsh Joshi and Mayank Kejriwal

Knowledge Enhanced Graph Neural Networks for Graph Completion (PDF)

Luisa Werner, Nabil Layaïda, Pierre Genevès and Sarah Chlyah

Lightweight Graph Convolutional Collaborative Filtering Recommendation Approach Incorporating Social Relationships (PDF)

Xiangfu Meng, Hongjin Huo, Xiaoyan Zhang and Wanchun Wang

Are GNNs the Right Tool to Mine the Blockchain? The Case of the Bitcoin Generator Scam (PDF)

Sam Yuen, Paula Branco, Aaron Chew, Guy-Vincent Jourdan, Fabian Lim and Laura Wynter

Tuesday 9:00 am – 10:30 am

Session 11: Society and Human (Applications III)
(*Room: Grace A*)
(Abstracts: Page 49)

Classification with Explanation for Human Trafficking Networks Detection (PDF)

Fabien Delorme, David Ing, Said Jabbour, Nelly Robin and Lakhdar Sais

Towards a (Semi-)Automatic Urban Planning Rule Identification in the French Language (PDF)

Maksim Koptelov, Margaux Holveck, Bruno Cremilleux, Justine Reynaud, Mathieu Roche and Maguelonne Teisseire

To Personalize or Not To Personalize? Soft Personalization and the Ethics of ML for Health (PDF)

Alessandro Falcetta, Massimo Pavan, Stefano Canali, Viola Schiaffonati and Manuel Roveri

MINDSET: A benchMarking suIte exploring seNsing Data for SElf sTates inference (PDF)

Christina Karagianni, Eva Paraschou, Sofia Yfantidou and Athena Vakali

Tuesday 9:00 am – 10:30 am

Session 12: Student Competition
(*Room: Grace D*)
(Abstracts: Page 50)

MAT: Effective Link Prediction via Mutual Attention Transformer (PDF)

Van Quan Nguyen, Quang Huy Pham, Quang Dan Tran, Kien Bao Thang Nguyen and Hieu Nghia Nguyen

Enhanced Edge Prediction, A Case Study: Predicting Links in Wikipedia Sites (PDF)

Apostolos Giannoulidis and Ioannis Mavroudopoulos

Link Prediction for Wikipedia Articles as a Natural Language Inference Task (PDF)

Chau Thang Phan, Quoc-Nam Nguyen and Kiet Nguyen

A Text-based Approach For Link Prediction on Wikipedia Articles (PDF)

Anh Tran, Tam Nguyen and Son Luu

Link Prediction on Graphs Using NLP Embedding (PDF)

João Victor Galvão da Mata and Martin Skovgaard Andersen

Predict Link Between Nodes Using An Ensemble Modelling Combining Depth Search Algorithm And Textual Similarity Score (PDF)

Aditya Kansal and Rishabh Mehta

Achieving High Performance in Link Prediction for Wikipedia Articles Using Ensemble Approach (PDF)

Weiwu Yang

Tuesday 10:30 am – 11:30 am

Keynote Talk: What Computers Know, and What They Should Know
(*Room: Grace C*)
By: Gerhard Weikum
Abstract: Page 30

Tuesday 11:30 am – 11:50 am

Coffee Break
(*Room: Grace B*)

Tuesday 11:50 am – 1:20 pm

Session 13: Feature and Label Learning (Research V)
(*Room: Grace C*)
(Abstracts: Page 52)

Sample Topology Exploration for Label Distribution Learning (PDF)

Yan-Wen Xiong, Heng-Ru Zhang, Fan Min and Peng-Cheng Li

Causal Feature Selection: Methods and a Novel Causal Metric Evaluation Framework (PDF)

Rezaur Rashid, Jawad Chowdhury and Gabriel Terejanu

ProPML: Probability Partial Multi-label Learning (PDF)

Łukasz Struski, Adam Pardyl, Jacek Tabor and Bartosz Zieliński

CaFe DBSCAN: A Density-based Clustering Algorithm for Causal Feature Learning (PDF)

Pascal Weber, Lukas Miklautz, Akshey Kumar, Claudia Plant and Moritz Grosse-Wentrup

Tuesday 11:50 am – 1:20 pm

Session 14: Science and Environment (Applications IV)
(*Room: Grace A*)
(Abstracts: Page 53)

Disaster Image Classification Using Pre-trained Transformer and Contrastive Learning Models (PDF)

Soudabeh Taghian Dinani and Doina Caragea

Non-Redundant Image Clustering of Early Medieval Glass Beads (PDF)

Lukas Miklautz, Andrii Shkabrii, Collin Leiber, Bendeguz Tobias, Benedict Seidl, Elisabeth Weissensteiner, Andreas Rausch, Christian Böhm and Claudia Plant

Exploring Deep Learning for Full-disk Solar Flare Prediction with Empirical Insights from Guided Grad-CAM Explanations (PDF)

Chetraj Pandey, Anli Ji, Trisha Nandakumar, Rafal Angryk and Berkay Aydin

Utilizing MODIS Fire Mask for Predicting Forest Fires Using Landsat-9/8 and Meteorological Data (PDF)

Yash Gupta, Navneet Goyal, Vishal John Varghese and Poonam Goyal

Tuesday 11:50 am – 1:20 pm

Session 15: Practical Applications of Explainable Artificial Intelligence Methods (PRAXAI I)
(*Room: Grace D*)
(Abstracts: Page 54)

Towards Explaining Satellite Based Poverty Predictions with Convolutional Neural Networks (PDF)

Hamid Sarmadi, Thorsteinn Rögnvaldsson, Mattias Ohlsson, Nils Roger Carlsson, Ibrahim Wahab and Ola Hall

Text Classification is Keyphrase Explainable! Exploring Local interpretability of Transformer Models with Keyphrase Extraction (PDF)

Dimitrios Akrivousis, Nikolaos Mylonas, Ioannis Mollas and Grigorios Tsoumakas

Interpreting Black-box Machine Learning Models for High Dimensional Datasets (PDF)

Md. Rezaul Karim, Md. Shajalal, Alex Graß, Till Döhmen, Sisay Adugna Chala, Christian Beecks and Stefan Decker

Enhanced Explanations for Knowledge-Augmented Clustering Using Subgroup Discovery (PDF)

Maciej Szelążek, Daniel Hudson, Szymon Bobek, Grzegorz J. Nalepa and Martin Atzmueller

Tuesday 1:20 pm – 2:30 pm

Lunch Break

Tuesday 2:30 pm – 4:00 pm

Tutorial: Interpretability Methods for Graph Neural Networks
(*Room: Grace C*)
By: Arijit Khan and Ehsan Bonabi Mobaraki
Abstract: Page 33

Tuesday 4:00 pm – 4:20 pm

Coffee Break
(*Room: Grace B*)

Tuesday 4:20 pm – 5:20 pm

Session 16: Medicine (Applications V)

(*Room: Grace C*)

(Abstracts: Page 55)

A Framework for Context-Sensitive Prediction in Time Series – Feasibility Study for Data-Driven Simulation in Medicine (PDF)

Fatoumata Dama, Christine Sinoquet and Corinne Lejus-Bourdeau

Optimizing Resource Allocation for Tumor Simulations over HPC Infrastructures (PDF)

Errikos Streviniotis, Nikos Giatrakos, Yannis Kotidis, Thalia Ntiniakou and Miguel Ponce de Leon

Death after Liver Transplantation: Mining Interpretable Risk Factors for Survival Prediction (PDF)

Veronica Guidetti, Giovanni Dolci, Erica Franceschini, Erica Bacca, Giulia Burastero, Davide Ferrari, Valentina Serra, Fabrizio Di Benedetto, Cristina Mussini and Federica Mandreoli

Tuesday 4:20 pm – 5:20 pm

Session 17: Journal I

(*Room: Grace A*)

(Abstracts: Page 56)

TOCOL: Improving Contextual Representation of Pre-trained Language Models via Token-Level Contrastive Learning (PDF)

Keheng Wang, Chuantao Yin, Rumei Li, Sirui Wang, Yunsen Xian, Wenge Rong and Zhang Xiong

GS2P: A Generative Pre-trained Learning to Rank Model with Over-parameterization for Web-Scale Search (PDF)

Yuchen Li, Haoyi Xiong, Linghe Kong, Jiang Bian, Shuaiqiang Wang, Guihai Chen and Dawei Yin

PANACEA: A Neural Model Ensemble for Cyber-Threat Detection (PDF)

Malik Al-Essa, Giuseppina Andresini, Annalisa Appice and Donato Malerba

Tuesday 4:20 pm – 5:20 pm

Session 18: Practical Applications of Explainable Artificial Intelligence Methods (PRAXAI II)

(*Room: Grace D*)

(Abstracts: Page 57)

Towards Quality Measures for xAI algorithms: Explanation Stability (PDF)

Marek Pawlicki

ORANGE: Opposite label soRting for tANGent Explanations in heterogeneous spaces (PDF)

Alejandro Kuratomi, Zed Lee, Ioanna Miliou, Tony Lindgren and Panagiotis Papapetrou

Instils Trust in Random Forest Predictions (PDF)

Gopal Jamnal and Gopal Jamnal

Tuesday 7:30 pm – 10:00 pm

Conference Dinner

Wednesday 11th of October 2023

Wednesday 9:00 am – 10:30 am

Session 19: Learning Methods and Theories (Research VI)
(*Room: Grace C*)
(Abstracts: Page 58)

Adaptive Clustered Federated Learning with Representation Similarity (PDF)

Chiyu Cai, Wei Wang and Yuan Jiang

Learning Representations through Contrastive Strategies for a more Robust Stance Detection (PDF)

Udhaya Kumar Rajendran, Amine Trabelsi and Amir Ben Khalifa

Toward a Realistic Benchmark for Out-of-Distribution Detection (PDF)

Pietro Recalcati, Fabio Garcea, Luca Piano, Fabrizio Lamberti and Lia Morra

On the Independence of Adversarial Transferability to Topological Changes (PDF)

Carina Newen and Emmanuel Müller

Wednesday 9:00 am – 10:30 am

Session 20: Industrial

(*Room: Grace A*)

(Abstracts: Page 59)

Prioritization of Identified Data Science Use Cases in Industrial Manufacturing via C-EDIF Scoring (PDF)

Raphael Fischer, Andreas Pauly, Rahel Wilking, Anoop Kini and David Graurock

Opportunistic Air Quality Monitoring and Forecasting with Expandable Graph Neural Networks (PDF)

Jingwei Zuo, Wenbin Li, Michele Baldo and Hakim Hacid

Short-term Forecast and Long-term Simulation for Accurate Energy Consumption Prediction (PDF)

Daniele Giampaoli, Francesca Cipollini, Denise Maffione and Luca Oneto

Practical Insights on Incremental Learning of New Human Physical Activity on the Edge (PDF)

Georgios Arvanitakis, Jingwei Zuo, Mthandazo Ndhlovu and Hakim Hacid

Wednesday 9:00 am – 10:30 am

Session 21: Journal II

(*Room: Grace D*)

(Abstracts: Page 61)

Hyperparameter Analysis of Wide-Kernel CNN Architectures in Industrial Fault Detection – An Exploratory Study (PDF)

Jurgen van den Hoogen, Dan Hudson, Stefan Bloemheuvel and Martin Atzmueller

Hybrid Approaches to Optimization and Machine Learning Methods (PDF)

Beatriz Flamia Azevedo, Ana Maria A. C. Rocha and Ana I. Pereira

Sparse Self-Attention Guided Generative Adversarial Networks for Time-Series Generation (PDF)

Nourhan Ahmed and Lars Schmidt-Thieme

Wednesday 10:30 am – 11:30 am

Keynote Talk: User-Defined Functions in Relational Databases:
Challenges and Promising Solutions based on YeSQL
(*Room: Grace C*)
By: Yannis Ioannidis
Abstract: Page 31

Wednesday 11:30 am – 11:50 am

Coffee Break
(*Room: Grace B*)

Wednesday 11:50 am – 1:20 pm

Session 22: Optimization (Research VII)
(*Room: Grace C*)
(Abstracts: Page 62)

AdaSub: Stochastic Optimization Using Second-Order Information in Low-Dimensional Subspaces (PDF)

João Victor Galvão da Mata and Martin Skovgaard Andersen

ISGP: Influence Maximization on Dynamic Social Networks Using Influence SubGraph Propagation (PDF)

Wan-Jhen Wu, Shiou-Chi Li and Jen-Wei Huang

HyperTab: Hypernetwork Approach for Deep Learning on Small Tabular Datasets (PDF)

Witold Wydmański, Oleksii Bulenok and Marek Śmieja

Wednesday 11:50 am – 1:20 pm

Session 23: Journal III
(*Room: Grace A*)
(Abstracts: Page 63)

DynamiSE: Dynamic Signed Network Embedding for Link Prediction (PDF)

Haiting Sun, Peng Tian, Yun Xiong, Yao Zhang, Yali Xiang, Xing Jia and Haofen Wang

PAF-Tracker: A Novel Pre-Frame Auxiliary and Fusion Visual Tracker (PDF)

Wei Liang, Derui Ding and Hui Yu

Entity Recognition Based on Heterogeneous Graph Reasoning of Visual Region and Text Candidate (PDF)

Xinzhi Wang, Nengjun Zhu, Jiahao Li, Yudong Chang and Zhennan Li

Wednesday 11:50 am – 1:20 pm

Session 24: Emerging Problems in Disinformation (DISA)
(*Room: Grace D*)
(Abstracts: Page 64)

Machine Learning-Based Android Malware Detection (PDF)

Carson Leung

Model Stitching Algorithm for Fake News Detection Problem (PDF)

Rafał Kozik, Aleksandra Pawlicka, Marek Pawlicki and Michal Choras

Towards Handling Bias in Intelligence Analysis with Twitter (PDF)

Alexandros Karakikes, Panagiotis Alexiadis, Theocharis Theocharopoulos, Nikolaos Skoulidas, Dimitris Spiliotopoulos and Konstantinos Kotis

A Continual Learning System with Self Domain Shift Adaptation for Fake News Detection (PDF)

Sebastián Basterrech, Andrzej Kasprzak, Jan Platos and Michal Wozniak

Combating Disinformation with Holistic Architecture, Neuro-symbolic AI and NLU Models (PDF)

Rafał Kozik, Wojciech Mazurczyk, Krzysztof Cabaj, Aleksandra Pawlicka, Marek Pawlicki and Michal Choras

Wednesday 1:20 pm – 2:30 pm

Lunch Break

Wednesday 2:30 pm – 4:00 pm

Session 25: Algorithms for Learning and Testing (Research VIII)
(*Room: Grace C*)
(Abstracts: Page 65)

Tackling Model Mismatch with Mixup Regulated Test-Time Training (PDF)

Bochao Zhang, Rui Shao, Jingda Du, Pc Yuen and Wei Luo

Natural Language Inference by Integrating Deep and Shallow Representations with Knowledge Distillation (PDF)

Pei-Chang Chen, Hao-Shang Ma and Jen-Wei Huang

Rapid and Scalable Bayesian AB Testing (PDF)

Srivas Chennu, Andrew Maher, Christian Pangerl, Subash Prabanantham, Jae Hyeon Bae, Jamie Martin and Bud Goswami

Finite-Sample Bounds for Two-Distribution Hypothesis Tests (PDF)

Cynthia Hom, William Yik and George Montanez

Wednesday 2:30 pm – 4:00 pm

Session 26: Computational Imaging, Vision, Linguistics and Language (CIVIL I)
(*Room: Grace A*)
(Abstracts: Page 67)

Adaptive Compressed Sensing for Real-Time Video Compression, Transmission, and Reconstruction (PDF)

Yaping Zhao, Qunsong Zeng and Edmund Lam

A CNN-Transformer Hybrid Network for Multi-scale object detection (PDF)

Jianhong Wu and Yingdong Ma

Searching Images in a Web Archive (PDF)

André Mourão and Daniel Gomes

ScaleFace: Uncertainty-aware Deep Metric Learning (PDF)

Roman Kail, Kirill Fedyanin, Nikita Muravev, Alexey Zaytsev and Maxim Panov

Wednesday 2:30 pm – 4:00 pm

Session 27: Smart City Data Analytics (SmartCities I)
(*Room: Grace D*)
(Abstracts: Page 68)

Empowering Urban Connectivity in Smart Cities using Federated Intrusion Detection (PDF)

Youcef Djenouri and Ahmed Nabil Belbachir

Recycling of Generic ImageNet-trained Models for Smart-city Applications (PDF)

Katarzyna Filus and Joanna Domanska

Incremental Targeted Mining in Sequences (PDF)

Kaixia Hu, Wensheng Gan, Gengsen Huang, Guoting Chen and Jerry Chun-Wei Lin

Price Prediction of Digital Currencies Using Machine Learning (PDF)

Ashutosh Dhar Dwivedi, Subhrangshu Adhikary, Subhayu Dutta and Jens Myrup Pedersen

ZigBee Network for AGV Communication in Industrial Environments (PDF)

Jarosław Flak, Tomasz Skowron, Rafał Cupek, Marcin Fojcik, Dariusz Caban and Adam Domański

Wednesday 4:00 pm – 4:20 pm

Coffee Break
(*Room: Grace B*)

Wednesday 4:20 pm – 5:20 pm

Session 28: Computational Imaging, Vision, Linguistics and Language (CIVIL II)
(*Room: Grace C*)
(Abstracts: Page 69)

YOLO-based Object Detection in Panoramic Images of Smart Buildings (PDF)

Sebastian Pokuciński and Dariusz Mrozek

Solving Inverse Problems in Compressive Imaging with Score-Based Generative Models (PDF)

Zhen Yuen Chong, Yaping Zhao, Zhongrui Wang and Edmund Lam

All Translation Tools Are Not Equal: Investigating the Quality of Language Translation for Forced Migration (PDF)

Ameeta Agrawal, Lisa Singh, Elizabeth Jacobs, Yaguang Liu, Gwyneth Dunlevy, Rhitabrat Pokharel and Varun Uppala

Wednesday 4:20 pm – 5:20 pm

Session 29: Graph Data Science and Applications (GraDSI I)
(*Room: Grace A*)
(Abstracts: Page 70)

JAMES: Normalizing Job Titles with Multi-Aspect Graph Embeddings and Reasoning (PDF)

Michiharu Yamashita, Jia Tracy Shen, Hamoon Ekhtiari, Thanh Tran and Dongwon Lee

Unfolding Temporal Networks through Statistically Significant Graph Evolution Rules (PDF)

Alessia Galdeman, Tommaso Locatelli, Matteo Zignani and Sabrina Gaito

Prediction of Future Nation-initiated Cyber Attacks from News-based Political Event Graph (PDF)

Bishal Lakha, Jason Duran, Edoardo Serra and Francesca Spezzano

Wednesday 4:20 pm – 5:20 pm

Session 30: Smart City Data Analytics (SmartCities II)
(*Room: Grace D*)
(Abstracts: Page 71)

Automated Detection of Trajectory Groups Based on SNN-Clustering and Relevant Frequent Itemsets (PDF)

Friedemann Schwenkreis

Object-aware Multi-criteria Decision-Making Approach Using Heuristic data-driven Theory for Intelligent Transportation Systems (PDF)

M S Mekala, Elyad Eyad and Gm Srivastava

Predicting Conflict Zones on Terrestrial Routes of Automated Guided Vehicles with Fuzzy Querying on Apache Kafka (PDF)

Bozena Malysiak-Mrozek, Stanisław Kozielski and Dariusz Mrozek

Low-Cost Gunshot Detection System with Localization for Community Based Violence Interruption (PDF)

Isaac Manring, James Hill, Paul Brantingham, George Mohler, Thomas Williams and Bruce White

Thursday 12th of October 2023

Thursday 9:00 am – 10:00 am

Session 31: Data Science for Social and Behavioral Analytics (DSSBA)
(*Room: Grace C*)
(Abstracts: Page 73)

Enhanced Mining of High Utility Patterns from Streams of Dynamic Profit (PDF)

Carson Leung

Towards Contiguous Sequences in Uncertain Data (PDF)

Zefeng Chen, Wensheng Gan, Gengsen Huang, Yanxin Zheng and Philip S. Yu

Emotion-based Dynamic Difficulty Adjustment in Video Games (PDF)

Krzysztof Kutt, Łukasz Ściga and Grzegorz J. Nalepa

Thursday 9:00 am – 10:00 am

Session 32: Graph Data Science and Applications (GraDSI II)
(*Room: Grace A*)
(Abstracts: Page 74)

Leveraging patient similarities via graph neural networks to predict phenotypes from temporal data (PDF)

Dimitrios Proios, Anthony Yazdani, Alban Bornet, Julien Ehrsam, Islem Rekik and Douglas Teodoro

Enhancing Recommendation Systems with Hybrid Manifold Regularized Knowledge Graph (PDF)

Giang Ngo and Nhi Vo

EvoAlign: A Continual Learning Framework for Aligning Evolving Networks (PDF)

Shruti Saxena and Joydeep Chandra

Thursday 9:00 am – 10:00 am

Session 33: Learning from Temporal Data (LfTD)
(*Room: Grace D*)
(Abstracts: Page 74)

1NN-DTW ARIMA LSTM: A New Ensemble for Forecasting Multi-domain/Multi-context Time Series (PDF)

Hadi Fanaee-T

Online Explainable Model Selection for Time Series Forecasting (PDF)

Amal Saadallah

LITE: Light Inception with boosTing tEchniques for Time Series Classification (PDF)

Ali Ismail-Fawaz, Maxime Devanne, Stefano Berretti, Jonathan Weber and Germain Forestier

Thursday 10:00 am – 11:00 am

Keynote Talk: Towards AI-Powered Data-Informed Educational Platforms

(*Room: Grace C*)

By: Sihem Amer-Yahia

Abstract: Page 32

Thursday 11:00 am – 11:20 am

Coffee Break

(*Room: Grace B*)

Thursday 11:20 am – 12:20 pm

Panel: Social Media and Misinformation

(*Room: Grace C*)

By: Charalampos Tsourakakis

Abstract: Page 34

Thursday 12:20 pm – 12:40 pm

Closing Ceremony

(*Room: Grace C*)

Thursday 12:40 pm – 1:40 pm

Lunch Break

Thursday 2:00 pm – 7:00 pm

Excursion: Trip to Vergina

Keynotes

Towards Quality-driven and AI-assisted Data Science

By **Angela Bonifati**
Professor of Computer Science
Lyon 1 University, France

Angela Bonifati is a Professor of Computer Science at Lyon 1 University and at the CNRS Liris research lab, where she leads the Database Group. She is also an Adjunct Professor at the University of Waterloo in Canada and a Senior member of French University Institute (IUF). Her current research interests are on several aspects of data management, including graph databases, knowledge graphs, data integration and data science. She has co-authored several publications in top venues of the data management field, including three Best Paper Awards, two books and an invited paper in ACM Sigmod Record 2018. She is the recipient of the TCDE Impact Award 2023 and a co-recipient of an ACM Research Highlights Award 2023. She was the Program Chair of ACM Sigmod 2022 and she is currently an Associate Editor for the Proceedings of VLDB and for several other journals, including the VLDB Journal, IEEE TKDE and ACM TODS. She is the President of the EDBT Executive Board (2020-2025) and was an interim member of the ACM Sigmod Executive Committee (2022-2023).

Abstract

One of the key processes of data science pipelines is data preparation, which aims at cleaning and curating the data for the subsequent analytical and inference steps. Data preparation deals with the errors and conflicts introduced into the input datasets during data collection and acquisition. These errors, such as violations of business rules, typos, missing values, replicated entries and abnormal features, are of different kinds depending on the nature of the data, ranging from structured data to graph-shaped data and time series. If these errors are kept into the data, they can propagate to the results of data science processes and also hamper their efficiency and trustworthiness.

The talk presents our latest results on enhancing the quality of querying and inference tasks in data science operating on different kinds of heterogeneous data. Among the others, we focus on real-life healthcare applications and provide the domain experts with useful AI-assisted data management techniques that can help them with their diagnoses and analyses. First, inconsistency-aware annotations can quantify the amount of quality for structured data input to analytical processes. These annotations are further exploited during query processing to enhance the output of queries with inconsistency degrees. Second, feature-based similarities among time series corresponding to patients' signals help to better identify groups of patients and to assess their risks for a particular disease. Third, violations of graph constraints can be addressed by human-guided feedback and lead to better accuracy of the repairing algorithms for graph-shaped data.

What Computers Know, and What They Should Know

By **Gerhard Weikum**
Scientific Director, Max Planck Institute for Informatics
Saarbruecken, Germany

Gerhard Weikum is a Scientific Director at the Max Planck Institute for Informatics in Saarbruecken, Germany, and an Adjunct Professor at Saarland University. He co-authored a comprehensive textbook on transactional systems, received the VLDB Test-of-Time Award 2002 for his work on automatic database tuning, and is one of the creators of the YAGO knowledge base which was recognized by the WWW Test-of-Time Award in 2018. Weikum is an ACM Fellow and elected member of various academies. He received the ACM SIGMOD Contributions Award in 2011, a Google Focused Research Award in 2011, an ERC Synergy Grant in 2014, the ACM SIGMOD Edgar F. Codd Innovations Award in 2016, and the Konrad Zuse Medal in 2021.

Abstract

Large knowledge graphs have become a key asset for search engines and other use cases. They are partly based on automatically extracting structured information from web contents and other texts, using a variety of pattern-matching and machine-learning methods. The semantically organized machine knowledge can be harnessed to better interpret text in news, social media and web tables, contributing to question answering, natural language processing (NLP) and data analytics.

A recent trend that has revolutionized NLP is to capture knowledge latently by billions of parameters of language models, learned at scale from huge text collections in a self-supervised manner. These pre-trained models form the basis of fine-tuning machine-learning solutions for tasks that involve both input texts and broad world knowledge, such as question answering, commonsense reasoning and human-computer conversations.

This talk reviews these advances and discusses lessons learned and limitations (see http://dx.doi.org/10.1561/1900000064 for a survey). Moreover, the talk identifies open challenges and new research opportunities. In particular, it discusses potential synergies of knowledge graphs and language models.

User-Defined Functions in Relational Databases: Challenges and Promising Solutions based on YeSQL

By **Yannis Ioannidis**

President of the Association of Computing Machinery (ACM)

Yannis Ioannidis is the President of the Association of Computing Machinery (ACM). He is a Professor at the Department of Informatics and Telecommunications of the University of Athens as well as an Associated Faculty at the "Athena" Research and Innovation Center, where he also served as the President and General Director for 10 years. His research interests include Database and Information Systems, Data Science, Data and Text Analytics, Data Infrastructures and Digital Repositories, Recommender Systems and Personalization, and Interactive Digital Storytelling. His work is often inspired by and applied to data management and analysis problems that arise in industrial environments or in the context of other scientific fields (Social Sciences and Humanities, Life Sciences, Physical Sciences) and the Arts. He is an ACM and IEEE Fellow, a member of Academia Europaea, and a recipient of several research, teaching, and service awards, including the Presidential Young Investigator Award in the US, the VLDB 10-Year Best Paper Award, and the ACM SIGMOD Contributions Award. He is currently the Greek delegate to the European Strategy Forum on Research Infrastructures (ESFRI) and a co-chair of the Global Climate Hub of the UN Sustainable Development Solutions Network.

Abstract

The diversity and complexity of modern data management applications have led to the extension of the relational paradigm with syntactic and semantic support for User-Defined Functions (UDFs). Although well-established in traditional DBMS settings, UDFs have become central in many application contexts as well, such as data science, data analytics, and edge computing. Still, a critical limitation of UDFs is the impedance mismatch between their evaluation and relational processing.

In this talk, I will first give an overview of the area and the technical challenges that UDF processing brings. I will then present YeSQL, an SQL extension with rich UDF support along with a pluggable architecture to easily integrate it with either server-based or embedded database engines. YeSQL currently supports Python UDFs fully integrated with relational queries as scalar, aggregator, or table functions. Key novel characteristics of YeSQL include easy implementation of complex algorithms and several performance enhancements, including tracing JIT compilation of Python UDFs, parallelism and fusion of UDFs, stateful UDFs, and seamless integration with a database engine. Experimental analysis showcases the usability and expressiveness of YeSQL and demonstrates that its techniques of minimizing context switching between the relational engine and the Python VM are very effective and achieve significant speedups of up to 68x in common, practical use cases compared to earlier approaches and alternative implementation choices.

Towards AI-Powered Data-Informed Educational Platforms

By **Sihem Amer-Yahia**
CNRS Research Director
Laboratoire d'Informatique de Grenoble, France

Sihem Amer-Yahia is a Silver Medal CNRS Research Director and Deputy Director of the Lab of Informatics of Grenoble. She works on exploratory data analysis and fairness in job marketplaces. Before joining CNRS, she was Principal Scientist at QCRI, Senior Scientist at Yahoo! Research and Member of Technical Staff at AT&T Labs. Sihem is PC chair for SIGMOD 2023 and vice president of the VLDB Endowment. She currently leads the Diversity, Equity and Inclusion initiative for the database community.

Abstract

The Covid-19 health crisis has seen an increase in the use of digital work platforms from videoconferencing systems to MOOC-type educational platforms and crowdsourcing and freelancing marketplaces. These levers for sharing knowledge and learning constitute the premises of the future of work. Educational technologies coupled with AI hold the promise of helping learners and teachers. However, they are still limited in terms of social interactions, user experience and learning opportunities as they must address a tension between learner-centered and platform-centered approaches. I will describe research at the intersection of data-informed recommendations and education theory and conclude with ethical considerations in building educational platforms.

Interpretability Methods for Graph Neural Networks

By **Arijit Khan** and **Ehsan Bonabi Mobaraki**

Abstract

The emerging graph neural network models (GNNs) have demonstrated great potential and success for downstream graph machine learning tasks, such as graph and node classification, link prediction, entity resolution, and question answering. However, neural networks are "black-box" – it is difficult to understand which aspects of the input data and the model guide the decisions of the network. Recently, several interpretability methods for GNNs have been developed, aiming at improving the model's transparency and fairness, thus making them trustworthy in decision-critical applications, leading to democratization of deep learning approaches and easing their adoptions. The tutorial is designed to offer an overview of the state-of-the-art interpretability techniques for graph neural networks, including their taxonomy, evaluation metrics, benchmarking study, and ground truth. In addition, the tutorial discusses open problems and important research directions.

Panel

Social Media and Misinformation

By **Charalampos Tsourakakis**
Boston University, United States

Dr. Charalampos Tsourakakis received his Ph.D. from the Algorithms, Combinatorics and Optimization (ACO) program at Carnegie Mellon University, and served as a Postdoctoral Fellow in Harvard University. He holds a Diploma in Electrical and Diploma Engineering from the National Technical University of Athens and a Master of Science from the Machine Learning Department at Carnegie Mellon University. Before joining Boston University, he worked as a researcher in the Google Brain team. He won a best paper award in IEEE Data Mining, has delivered three tutorials in the ACM SIGKDD Conference on Knowledge Discovery and Data Mining, and has designed two graph mining libraries for large-scale graph mining, one of which has been officially included in Windows Azure. His research focuses on large-scale graph mining, and machine learning.

Abstract

Misinformation on social media has significant, and sometimes dramatic, consequences, causing harm to individuals, communities, and even entire societies. We have seen how the dissemination of disinformation on social media affects everything from vaccines or the treatment of diseases to the spread of false information in an attempt to destabilize our governments or influence our votes. The rapid spread of misinformation on social media can make it difficult for people to distinguish fact from fiction and can undermine trust in reliable sources of information.

The DSAA panel will discuss how computational methods such as natural language processing, machine learning, and network analysis can be used to detect and address misinformation on social media. These methods can help identify patterns and trends in the spread of misinformation and can be used to develop targeted interventions to counter it. For example, AI algorithms can be used to analyze the language and content of social media posts to detect false or misleading information, while network analysis can help identify the sources of misinformation and the individuals or groups who are most vulnerable to its effects. By leveraging these computational methods, we can better understand the dangers of misinformation on social media and develop effective strategies to mitigate its impact.

Participants

- Ramon Salaverria Aliaga (University of Navarra),
- David Camacho (Polytechnic University of Madrid),
- Ioannis Kompatsiaris (ITI/CERTH)
- Paolo Rosso (Polytechnic University of Valencia)

Papers' Abstracts

The Proceedings are indexed by IEEE Xplore with the following ISBN:
979-8-3503-4503-2

Session: Advanced Analytics and Knowledge Discovery Methods (Research I)

Designing Concept Drift Detection Ensembles: A Survey (PDF)

Martin Trat and Jivka Ovtcharova

Abstract:

Data streams represent real-world concepts that typically cannot be assumed to be static. Instead, sensors monitoring industrial resources deteriorate, network intrusion attacks exhibit new patterns, consumer behavior or public interest in news topics change. These are examples of a phenomenon commonly known as concept drift. It threatens the performance of estimation models conducting inference on associated data. Explicitly employed concept drift detection is among the most promising approaches to maintain a stable and robust performance of productively used estimators, as it identifies the times when adaptation to new concepts becomes actually necessary. Combining concept drift detectors to form ensembles can additionally increase detection performance as it aggregates their individual specializations and allows advanced processing of concept drift evidence. This survey is the first to systematically retrieve and overview research relevant in the field of concept drift detection ensembles. It identifies the research as being in its initial stages and reveals numerous gaps, which should be addressed to further exploit the substantial performance potential ensembles have over individually applied detectors. Furthermore, it contributes to the literature by discussing and describing key characteristics and principles of their design in a structured manner.

Sliding Window Sampling over Data Stream — A Solution Based on Devil's Staircases (PDF)

Dominik Bojko, Jacek Cichoń and Mirosław Kutyłowski

Abstract:

The paper concerns sampling from a data stream {Si}: at a moment t the sampler should hold a value St−j, where j ∈ {0,...,n−1} should be chosen according to an a priori specified probability distribution D on {0,..., n−1}, where D as well as the *window size* n are fixed and do not depend on t. We assume that the sampler has a constant size memory, while n might be large, so the sampler cannot remember the last n values of the stream except for a few. The problem is that the window of the last n elements changes at each step and when we have to resample, then almost all values from which we have to choose are already forgotten. The case of uniform distribution D has been considered by Braverman, Ostrovsky, and Zaniolo in 2013. We present an alternative generic approach based on specific Markov chains called *devil's staircases*. Unlike the previous solution, it is not limited to the uniform distribution: it generates a sample according to any admissible distribution in the window of size n and uses memory of size O(1). We provide sufficient conditions for the distribution D to be admissible. Although the class of such distributions is quite wide from the point of view of practical applications, we show some natural limitations for this class.

slidSHAPs: sliding Shapley Values for correlation-based change detection in time series (PDF)

Chiara Balestra, Bin Li and Emmanuel Müller

Abstract:

For volatile multivariate time series, variations in the distributions of the input dimension and the correlation structure present an open challenge. The different distributions before and after a change-point hinder the performance of most of the predictive methods, mostly requiring re-training of the models. The detection of such change points represents a severe problem, as volatile data labeling is often either expensive or delayed in streaming data; Moreover, classical concept drift detectors usually struggle with detecting changes in correlations of multivariate time series' input variables. We focus on unsupervised change detection, tracking correlation changes in the input variables without class labels. By introducing slidSHAPs, we propose a fully unsupervised change detector for multivariate time series with categorical value domains; our tool detects correlation-based changes through a representation of the correlation structure of the input data. The slidSHAPs series underlines distributional changes even in a few univariate input variables, thus, being more sensitive to changes than any prior change point detection method. In contrast to the well-known application of Shapley values for interpretable machine learning, we use this foundational game-theoretic concept to extrapolate information on the correlation structure of data streams and achieve higher sensitivity towards multiple changes in the empirical evaluation of synthetic and real-world data.

Measurement of Illegal Android Gambling App Ecosystem From Joint Promotion Perspective (PDF)

Yadi Han, Shanshan Wang, Yiwen Li, Xueyang Cao, Limei Huang and Zhenxiang Chen

Abstract:

With the continuous intensification of China's crack-down on gambling and the rapid development of Internet technology, offline gambling has gradually shifted to online, and operations have been extended from domestic to overseas. Confronted with strict censorship regulations, illegal online gambling websites can employ dark links and black hat search engine optimization (SEO) approaches to promote and distribute mobile gambling apps(applications). To expose China's gambling application ecosystem, the researchers conducted a cluster analysis of illegal gambling apps. However, they did not consider the interdependence between their distribution channels and upstream promotion websites. For the first time, we analyze and measure the gambling application ecosystem at the promotion community level in conjunction with the promotion relationship between the gambling apps and their upstream gambling pages. The clustering relationships of gambling apps, domains and IPs, and payment platforms in gambling communities with similar promotion preferences are mainly analyzed. From 277,816 related domains of suspected illegal industries, these illegal gambling websites' upstream and downstream relationships are screened and separated in detail, forming a knowledge graph of online gambling apps centered on promotion relationships. Based on this graph, information integration of entities with different functions in the gambling industry chain (promotion, gambling pages, mobile apps) is carried out. This research provides a new perspective on the profit chain research of illegal online apps in conjunction with the promotion relationship of gambling pages. We have also published this dataset to support further in-depth analysis and black industry governance work in the security community at https://github.com/yadispace/gambling app KG.

Session: Business and Industry (Applications I)

Electricity Price Forecasting Based on Order Books: A Differentiable Optimization Approach (PDF)

Léonard Tschora, Tias Guns, Erwan Pierre, Marc Plantevit and Celine Robardet

Abstract:

We consider day-ahead electricity price forecasting on the European market. In this market, participants can offer electricity for sale or purchase for a specific price by submitting overnight orders. Market operators determine the market clearing price – the price at which the amount of electricity supplied equals the amount of electricity demanded – using the Euphemia balancing algorithm. EUPHEMIA is a quadratic optimization problem that maximizes the social welfare defined as the sum of the supplier surplus and consumer surplus while ensuring a null energy balance. This mechanism deeply influences the price calculation, but has so far been little considered in electricity price forecasting algorithms. Existing models are generally based on identifying relationships between exogenous characteristics (consumption and production forecasts) and the market clearing price to be predicted. A few studies have examined the EUPHEMIA mechanism during prediction, by doing costly manual transformations on order books. In this article, we overcome this limitation by considering the pricing mechanism during model training. For this, we use a predict-and-optimize strategy with differentiable optimization. We design a fully differentiable and scalable solving method for the EUPHEMIA optimization problem and apply it on real-life data from the European Power Exchange (EPEX). We design different model architectures using our differentiable solver and empirically study the impact of taking into account the optimal calculation of prices within the training of the neural network.

Contextual Advertising Strategy Generation via Attention and Interaction Guidance (PDF)

Issam Benamara and Emmanuel Viennet

Abstract:

Digital advertising has become one of the most important aspects of modern marketing, as it allows businesses to reach larger audiences with greater precision, controllable cost, and measurable feedback. However, the process of designing effective advertising strategies relies heavily on human experts and thus remains suboptimal. In this paper, we propose a novel method for Contextual Advertising Strategy Generation via

Attention and InteRaction Guidance (ASGAR) that leverages transformers and a soft contrastive learning approach to optimize campaign performance. An advertising strategy is a combination of multiple targeting options, and its performance is tied strictly to the combination as a whole. This makes the exploration of the high combinatorial space infeasible and autoregressive methods inefficient. Therefore, constraints of non-combinatorial exploration and non-autoregressive generation have to be met. To the best of our knowledge, this is the first method that satisfies all constraints while also outperforming the previous methods. We compare our results with state-of-the-art methods on a public data set and with human experts in a company-deployed environment. We show that our method can effectively generate high-performance advertising strategies with better stability and controllable exploration.

HRGCN: Heterogeneous Graph-level Anomaly Detection with Hierarchical Relation-augmented Graph NeuralNetworks (PDF)

Jiaxi Li, Guansong Pang, Ling Chen and Mohammad-Reza Namazi-Rad

Abstract:

This work considers the problem of heterogeneous graph-level anomaly detection. Heterogeneous graphs are commonly used to represent behaviours between different types of entities in complex industrial systems for capturing as much information about the system operations as possible. Detecting anomalous heterogeneous graphs from a large set of system behaviour graphs is crucial for many real-world applications like online web/mobile service and cloud access control. To address the problem, we propose HRGCN, an unsupervised deep heterogeneous graph neural network, to model complex heterogeneous relations between different entities in the system for effectively identifying these anomalous behaviour graphs. HRGCN trains a hierarchical relation-augmented Heterogeneous Graph Neural Network (HetGNN), which learns better graph representations by modelling the interactions among all the system entities and considering both source-to-destination entity (node) types and their relation (edge) types. Extensive evaluation on two real-world application datasets shows that HRGCN outperforms state-of-the-art competing anomaly detection approaches. We further present a real-world industrial case study to justify the effectiveness of HRGCN in detecting anomalous (e.g., congested) network devices in a mobile communication service. HRGCN is available at https://github.com/jiaxililearn/HRGCN.

FIW-GNN: A Heterogeneous Graph-based Learning Model for Credit Card Fraud Detection (PDF)

Kuan Yan, Junbin Gao and Dmytro Matsypura

Abstract:

The global economic losses caused by credit card fraud are enormous and continuously increasing. Effective and accurate fraud detection has become a crucial task in recent years. Prior approaches can achieve good detection performance under certain conditions. However, these existing methods lack the robustness and scalability to deal with real-world credit card transactional datasets containing a large number of missing values. In this paper, we propose a Feature Importance-based Weighted Graph Neural Network (FIW-GNN) as an effective, stable, and practical solution for credit card fraud detection. First, we propose a method to construct a heterogeneous graph designed for credit card transactional datasets. Next, based on the architecture of the relational graph convolutional network, a feature importance-based method is employed to assign edge weights. Finally, we evaluate the effectiveness of FIW-GNN on two benchmark datasets. The experimental results demonstrate that FIW-GNN outperforms the state-of-the-art baselines in all selected evaluation metrics.

Session: Private, Secure, and Trust Data Analytics (PSTDA I)

Stochastic Perturbation Averaging Boosts Transferability of Adversarial Examples (PDF)

Rongbo Yang, Qianmu Li and Shunmei Meng

Abstract:

In image, video and even real physics domains, adversarial examples can mislead deep models to produce wrong predictions. Transfer-based attacks against black-box models are more in line with realistic scenarios, but adversarial examples made on surrogate model have a low success rate when transferred to the target model due to overfitting the source model. We study the Stochastic Weight Averaging strategy in the domain generalization process and propose a Stochastic Perturbation Averaging method (SPA). Specifically, we add stochastic perturbations to the examples during the gradient descent attack, and we design a Central

Amplification method (CAM) to enhance this random variation, then SPA stabilizes the iteration direction by computing the gradient average of the perturbed examples to find a relatively flat local minimum of the loss function. SPA is an efficient and general strategy, which can significantly improve the transferability of the gradient-based attack methods. For instance, the average attack success rate of the adversarial examples produced based on four single models against seven pre-trained models reached 90.10%, which is the best result so far. Code is available at https://github.com/yangrongbo/SPA.

Novel Few-shot Learning Based Fuzzy Feature Detection Algorithms (PDF)

Yun Luo, Liangfu Lu, Xudong Cui, Yan Du, Yingying Bi, Limin Zhu and Christy Jie Liang

Abstract:

The Internet of Things (IoT) has significantly enhanced various aspects of our daily lives, including security, health, education, and energy efficiency, among others. Within the realm of IoT, image classification stands as a pivotal technique that has achieved notable success in domains such as facial recognition within security and scene recognition in transportation for traffic analysis. Nonetheless, the challenge emerges when tackling classification tasks with only limited labeled samples available for each category. Conventional machine learning techniques often struggle to attain satisfactory classification results under such circumstances. To address this issue, the concept of few-shot learning has emerged, aiming to achieve effective classification using only a small number of labeled samples. State-of-the-art few-shot learning models have introduced novel frameworks to tackle this problem. However, the inherent ambiguity and uncertainty within data often hinder the performance of classification methods. To overcome this limitation, this paper proposes the integration of fuzzy learning with few-shot learning in the context of feature extraction. The objective is to mitigate data fuzziness and enhance model performance. Leveraging a fuzzy extraction algorithm, we introduce fuzzy prototype networks and a fuzzy graph neural network with fuzzy reasoning. These frameworks are designed to analyze noisy and uncertain data, utilizing convolutional neural networks for feature extraction and applying fuzzy reasoning to capture ambiguity representations for features within each fuzzy set. The SoftMax function is then normalized to serve as a feature weight, effectively constraining the original feature vector. The effectiveness and efficiency of our proposed model are demonstrated through experimental evaluations conducted on various public datasets. The results showcase the model's capability in addressing the challenges posed by limited labeled data and data uncertainty, thus reaffirming its potential in enhancing the performance of image classification tasks within the IoT context.

A Contextualized Transformer-Based Method for Cyberbullying Detection (PDF)

Nabi Rezvani, Amin Beheshti and Xuyun Zhang

Abstract:

Automatic detection of Cyberbullying is a challenging task due to the availability of limited trained data, which is usually noisy and inherently multimodal. Transfer learning over pre-trained BERT-based language models has succeeded in various complex use cases like sequence-to-sequence translation and text classification. These methods mainly utilize transformer models to learn the word and sentence-level relationships. While they have demonstrated promising results, they only focus on textual features without taking contextual and structural information into account. Moreover, due to the data-heavy nature of BERT-based models, they may fail to model all the desired relationships if not adequate training data is provided to them during the fine-tuning process. In this paper, we propose a novel Session-level Contextualized Transformer-based architecture for Cyberbullying Detection (SECTR-CD), which can leverage transfer learning for modeling word-level attention while also being able to model sentence-level relationships in large bodies of text. The model is also capable of utilizing other contextual features from various modalities like images and social information. Our experimental results indicate remarkable improvement in the Cyberbullying detection task even in the presence of limited training samples.

Privacy-Preserving Learning via Data and Knowledge Distillation (PDF)

Fahim Faisal, Carson K. Leung, Noman Mohammed and Yang Wang

Abstract:

In the current era of data science, deep learning, computer vision and image analysis have become ubiquitous across various sectors, ranging from government agencies and large corporations to small end devices, due to their ability to simplify people's lives. However, the widespread use of sensitive image data and the high memorization capacity of deep learning present significant privacy risks. Now, a simple Google search can yield numerous images of a person, and the knowledge that a specific patient's record was utilized for training a

specific model associated with a disease may reveal the patient's ailment, potentially leading to membership privacy leakage and other advanced attacks in the future. Furthermore, these unprotected models may also suffer from poor generalization due to this overfitting to train data. Previous state-of-the-art methods like differential privacy (DP) and regularizer-based defenses compromised functionality, i.e., task accuracy, to preserve privacy. Such an imbalanced trade-off raises concerns about the practicability of such defenses. Other existing knowledge-transfer-based methods either reuse private data or require more public data, which could compromise privacy and may not be viable in certain domains. To address these challenges, where membership privacy is of utmost importance and utility cannot be compromised, we propose a novel collaborative distillation approach that transfers the private model's knowledge based on a minimal amount of distilled synthetic data, leading to a compact private model in an end-to-end fashion. Empirically, our proposed method guarantees superior performance compared to most advanced models currently in use, increasing utility by almost 8%, 34%, and 6% for CIFAR-10, CIFAR-100, and MNIST, respectively. The utility resembles non-private counterparts almost closely while maintaining a respectable level of membership privacy leakage of 50-53.5%, despite employing a smaller model with 50% fewer parameters.

Session: Advanced Classification Methods (Research II)

Evaluating Explanation Methods of Multivariate Time Series Classification through Causal Lenses (PDF)

Etienne Vareille, Adel Abbas, Michele Linardi and Vassilis Chrsitopides

Abstract:

Explainable machine learning techniques (XAI) aim to provide a solid descriptive approach to Deep Neural Networks (NN). In Multi-Variate Time Series (MTS) analysis, the most recurrent techniques use relevance attribution, where importance scores are assigned to each TS variable over time according to their importance in classification or forecasting. Despite their popularity, post-hoc explanation methods do not account for causal relationships between the model outcome and its predictors. In our work, we conduct a thorough empirical evaluation of model-agnostic and model-specific relevance attribution methods proposed for TCNN, LSTM, and Transformers classification models of MTS. The contribution of our empirical study is three-fold: (i) evaluate the capability of existing post-hoc methods to provide consistent explanations for high-dimensional MTS (ii) quantify how post-hoc explanations are related to sufficient explanations (i.e., the direct causes of the target TS variable) underlying the datasets, and (iii) rank the performance of surrogate models built over post-hoc and causal explanations w.r.t. the full MTS models. To the best of our knowledge, this is the first work that evaluates the reliability and effectiveness of existing xAI methods from a temporal causal model perspective.

Interpretable Time Series representation for Classification Purposes (PDF)

Etienne Le Naour, Ghislain Agoua, Nicolas Baskiotis and Vincent Guigue

Abstract:

Deep learning has made significant advances in creating efficient representations of time series data by automatically identifying complex patterns. However, these approaches lack interpretability, as the time series is transformed into a latent vector that is not easily interpretable. On the other hand, Symbolic Aggregate approximation (SAX) methods allow the creation of symbolic representations that can be interpreted but do not capture complex patterns effectively. In this work, we propose a set of requirements for a neural representation of univariate time series to be interpretable. We propose a new unsupervised neural architecture that meets these requirements. The proposed model produces consistent, discrete, interpretable, and visualizable representations. The model is learned independently of any downstream tasks in an unsupervised setting to ensure robustness. As a demonstration of the effectiveness of the proposed model, we propose experiments on classification tasks using UCR archive datasets. The obtained results are extensively compared to other interpretable models and state-of-the-art neural representation learning models. The experiments show that the proposed model yields, on average better results than other interpretable approaches on multiple datasets. We also present qualitative experiments to assess the interpretability of the approach.

LSFuseNet: Dual-Fusion of Landsat-8 and Sentinel-2 Multispectral Time Series for Permutation Invariant Applications (PDF)

Arshveer Kaur, Poonam Goyal and Navneet Goyal

Abstract:

Satellite data provides valuable insights into environmental changes and natural resource management, such as monitoring deforestation, mapping land use changes, and identifying areas at risk of soil degradation. Landsat-8 and Sentinel-2 are the publicly available high spatial resolution satellites launched in recent years. But, both have a moderate temporal resolution which limits their use in the applications like precision agriculture, land cover mapping, disaster monitoring, etc. For such applications, daily or weekly monitoring is better suited. Fusing data from the two satellites can provide enhanced observations. Both Landsat-8 and Sentinel-2 satellites have the same geographic coordinate systems, which makes them amiable for fusion. But, fusing data at the pixel level for these satellites is challenging as they visit the same location on different days. The proposed model 'LSFuseNet' effectively fuses data at the feature level. It is a dual-fusion model in which bi-directional cross-modal attention is used to identify and exchange the hot-spot information in the two modalities. A feature alignment module learns the fine-grained features and mitigates the noise in the data. We have innovatively applied contrastive learning to improve the quality of the learned representations of the data from the two satellites. We evaluate our model for two applications – crop yield prediction and snow cover prediction. For crop yield prediction, we have taken two crops, viz. corn, and soybean, for approximately 500 counties in the US. For snow cover prediction, we considered approximately 1300 US counties. Our extensive experiments show that LSFuseNet outperforms competing models. Also, the benefit of fusing the data from two satellites over using the data from a single satellite is evident from the results of both applications. We have further modified the model to include meteorological and/or soil data (if applicable) to further enhance the performance of the model.

A Novel Method for Temporal Graph Classification Based on Transitive Reduction (PDF)

Carolina Jeronimo, Zenilton Patrocínio Jr., Simon Malinowski, Guillaume Gravier and Silvio Guimaraes

Abstract:

Domains such as bio-informatics, social network analysis, and computer vision, describe relations between entities and cannot be interpreted as vectors or fixed grids, instead, they are naturally represented by graphs. Often this kind of data evolves over time in a dynamic world, respecting a temporal order being known as temporal graphs. The latter became a challenge since subgraph patterns are very difficult to find and the distance between those patterns may change irregularly over time. While state-of-the-art methods are primarily designed for static graphs and may not capture temporal information, recent works have proposed mapping temporal graphs to static graphs to allow for the use of conventional static kernels and graph neural approaches. In this study, we compare the transitive reduction impact on these mappings in terms of accuracy and computational efficiency across different classification tasks. Furthermore, we introduce a novel mapping method using a transitive reduction approach that outperforms existing techniques in terms of classification accuracy. Our experimental results demonstrate the effectiveness of the proposed mapping method in improving the accuracy of supervised classification for temporal graphs while maintaining reasonable computational efficiency.

Session: Business and Education (Applications II)

Enhancing the Performance of Automated Grade Prediction in MOOC using Graph Representation Learning (PDF)

Soheila Farokhi, Aswani Yaramala, Jiangtao Huang, Muhammad Fawad Akbar Khan, Xiaojun Qi and Hamid Karimi

Abstract:

In recent years, Massive Open Online Courses (MOOCs) have gained significant traction as a rapidly growing phenomenon in online learning. Unlike traditional classrooms, MOOCs offer a unique opportunity to cater to a diverse audience from different backgrounds and geographical locations. Renowned universities and MOOC-specific providers, such as Coursera, offer MOOC courses on various subjects. Automated assessment tasks like grade and early dropout predictions are necessary due to the high enrollment and limited direct interaction between teachers and learners. However, current automated assessment approaches overlook the structural links between different entities involved in the downstream tasks, such as the students and courses. Our hypothesis suggests that these structural relationships, manifested through an interaction graph, contain valuable information that can enhance the performance of the task at hand. To validate this, we construct a unique knowledge graph for a large MOOC dataset, which will be publicly available to the research community.

Furthermore, we utilize graph embedding techniques to extract latent structural information encoded in the interactions between entities in the dataset. These techniques do not require ground truth labels and can be utilized for various tasks. Finally, by combining entity-specific features, behavioral features, and extracted structural features, we enhance the performance of predictive machine learning models in student assignment grade prediction. Our experiments demonstrate that structural features can significantly improve the predictive performance of downstream assessment tasks. The code and data are available in https://github.com/DSAatUSU/MOOPer grade prediction.

Supplier Qualification Document Recognition through Open-set Recognition (PDF)

Giuseppe Rizzo and Angelo Impedovo

Abstract:

Large and medium-sized manufacturing companies are concerned with maintaining their supplier registries with well-reputed suppliers sourced over time. Every supplier periodically undergoes rigorous qualification processes where procurement officers assess, among other factors, the supplier document compliance status. To this end, procurement officers periodically ask suppliers, via digital e-procurement platforms, for qualification documents of different categories. Conversely, suppliers promptly answer by handing out such documents that can, maliciously or inadvertently, be wrong. When wrong qualification documents remain undetected, and the associated suppliers are qualified, a threat to the overall business arises: procurement officers may entrust purchase orders to not compliant suppliers with unpredictable performances. Our claim is that equipping e-procurement platforms with document recognition based on supervised open-set recognition (OSR) could mitigate the problem. In particular, we deem OSR solutions suitable for supplier qualification document recognition due to their simultaneous abilities of i) recognizing documents belonging to relevant categories and ii) rejecting those belonging to unknown categories. Quantitative and qualitative results from a real-world case study in partnership with an Italian manufacturing company show that the proposed solution is viable.

Identifying Survival-Changing Sequential Patterns for Employee Attrition Analysis (PDF)

Youssef Oubelmouh, Frédéric Fargon, Cyril de Runz, Arnaud Soulet and Cyril Veillon

Abstract:

Employee attrition is a pervasive problem for many organizations, and reducing it has become a key goal in the business world. Although there is a substantial body of literature on predicting customer attrition, the literature on employee attrition is comparatively limited. Moreover, even studies that do address employee attrition often fail to consider the impact of time and duration on attrition rates. In this context, the present paper aims to fill this gap in the literature by combining frequent pattern mining in sequences of events and survival analysis with Kaplan-Meier to examine how event sequences affect employee attrition. We introduce the notion of survival-changing sequential patterns that highlight events that significantly impact the survival estimator. Our findings suggest that certain patterns are associated with a higher rate of employee retention, while the addition of specific events can have a positive or negative impact on employee survival. This research highlights the importance of analyzing event sequences and duration when attempting to reduce employee attrition rates. The practical implications of this research are significant, as it provides a framework for organizations seeking to retain their employees and enhance their overall performance.

Towards Deep Learning Models for Automatic Computer Program Grading (PDF)

Peter Nagy and Heidar Davoudi

Abstract:

Automatic grading of computer programs has a great impact on both computer science education and the software industry as it saves human evaluators a tremendous amount of time required for assessing programs. However, to date, this problem lacks extensive research from the machine learning/deep learning perspective. Currently, the traditional auto-grading systems are mostly based on test-case execution results. However, these approaches lack insight into the syntax and semantics of the codes, and therefore, are far from human-level evaluation. In this study, we leverage the power of language models pre-trained on programming languages. We introduce two simple deep architectures and show that they consistently outperform the shallow models built upon extensive feature engineering approaches by a high margin. We also develop an incremental transductive learning algorithm that only requires a single reference solution to a problem and takes advantage of the correct implementations in the set of programs to be evaluated. Furthermore, our human evaluation results show that the proposed approaches provide partial marks having a strong correlation with marks given

by human graders. We prepare and share a dataset of C++ and Python programs for future research (Code and data are available at https://github.com/peter-nagy1/Deep-Grader).

Session: Private, Secure, and Trust Data Analytics (PSTDA II)

Privacy-aware Adaptive Collaborative Learning Approach for Distributed Edge Networks (PDF)

Saeed Alqubaisi, Deepak Puthal, Joy Dutta and Ernesto Damiani

Abstract:

To facilitate the Edge AI paradigm in distributed networks, we propose novel collaborative learning methodologies for a connected network of edge nodes. Our proposed methodologies tackle the challenges in distributed learning where there are constraints on data privacy and a low degree of overlap between the classes observed by the nodes. These approaches entail sharing class distribution information between nodes, computing nodes, and class weights, training local models on each node, then aggregating the models using the determined weights. It favors nodes that have encountered unique or less common classes in their local datasets. Through a series of experiments using an activity recognition dataset, we demonstrate the effectiveness and scalability of our proposed approaches. We show the adaptive nature of the proposed approach by achieving classification accuracy above the baseline, even with little overlap between the observed classes. This study serves as a foundation for future advancements in collaborative learning on edge networks, and encourages the development of scalable solutions.

Multi-Granularity Entity Recognition Based Sentence Ranking for Multi-Document Summarization (PDF)

Guowei Zhang, Xuyun Zhang, Zhiyong Wang and Amin Beheshti

Abstract:

Text summarization aims to condense text documents into a concise textual summary, which improves the efficiency of people in comprehending information. While deep learning-based summarization methods for individual documents have achieved good performance, there is an increasing demand for summarizing multiple related documents of a topic or event can yield a more coherent and succinct summary of the document set. However, the characteristics of multiple documents with more information, longer texts, and different styles impose new challenges to existing methods in dealing with the multi-aspect of a topic or an event. Therefore, in this paper, we propose a novel multi-granularity model with entity recognition for better sentence ranking and capturing the key information of different documents with a comprehensive and accurate summary. Specifically, we use PRIMERA as a token encoder based on the encoder-decoder framework. Then, a named entity recognition model is trained to identify key elements in documents such as people, location, organization, etc. The proposed model will focus more on these key elements. Based on the named entity recognition results, we further devise a sentence ranking module that allows the model to assign different weights to different sentences based on the sum of the frequencies of the entities contained in the sentences. Finally, based on the results of the multi-granularity encoding vector, the decoder can generate a comprehensive and accurate summary. To evaluate the performance of our proposed model, we conducted experiments on CoNLL2003, DUC2003, and DUC2004, which demonstrated the performance improvement of our proposed method over four previous models.

Temporal Differential Privacy for Human Activity Recognition (PDF)

Debaditya Roy and Sarunas Girdzijauskas

Abstract:

Differential privacy (DP) is a method to protect individual privacy when the data is used for downstream analytical tasks. The core ability of DP to quantify privacy numerically separates it from other privacy-preserving methods. In human activity recognition (HAR), differential privacy can protect users' privacy who contribute their data to train machine learning algorithms. While some methods are developed for privacy protection in such cases, no method quantifies privacy and seamlessly integrates into machine learning frameworks like DP. The paper proposes a DP framework called *TEMPDIFF* (short for temporal differential privacy), which guarantees privacy preserving human activity recognition for wearable time-series data with competitive classification performance and works with any machine-learning/deep-learning methods. *TEMPDIFF* capitalizes on the temporal characteristics of wearable sensor data to improve the modelling task, which enhances the privacy-utility tradeoff. *TEMPDIFF* uses ensembling and a novel *temporal partitioning* algorithm for time-series

data to ensure optimal training of ensemble models. In *TEMPDIFF*, consensus through ensembling and the addition of controlled Laplacian noise obscures sensitive information used to train the models, guaranteeing strict levels of differential privacy. The proposed method is evaluated on two popular HAR datasets. It outperforms the classification accuracy and privacy budget for both datasets compared to the state-of-the-art approaches.

Graph Disentangled Collaborative Filtering Based on Multi-order Similarity Constraint (PDF)

Yaoze Liu, Junwei Du, Haojie Li and Guanfeng Liu

Abstract:

Disentangled collaborative filtering can explicitly generate embeddings based on users' interests and help improve the interpretability and robustness of recommendations. However, the existing disentangled graph collaborative filtering methods rely solely on direct interaction constraints between nodes to learn node embeddings, which cannot represent higher-order constraints between nodes and node-type differences, resulting in suboptimal node representations and negatively affecting recommendation performance. To address this problem, we propose a Multi-order Similarity Constraint Disentangled Graph Collaborative Filtering (DGCF-MSC) method, which considers not only direct interaction constraints between nodes but also designs a neighborhood enhancement mechanism based on high-order relationships between homogeneous nodes. We realize the disentanglement of heterogeneous type nodes in different feature spaces in a graph convolutional neural network to make the generated embedding more interpretable and improve the performance of graph collaborative filtering. We conduct extensive experiments with three recommendation system datasets and the results demonstrate that DGCF-MSC outperforms the existing disentangled graph collaborative filtering methods in all performance metrics. Our code is released on https://github.com/lustrelake/DGCF_MSC.

Session: Time Series and Forecasting (Research III)

Combining Forecasts using Meta-Learning: A Comparative Study for Complex Seasonality (PDF)

Grzegorz Dudek

Abstract:

In this paper, we investigate meta-learning for combining forecasts generated by models of different types. While typical approaches for combining forecasts involve simple averaging, machine learning techniques enable more sophisticated methods of combining through meta-learning, leading to improved forecasting accuracy. We use linear regression, k-nearest neighbors, multilayer perceptron, random forest, and long short-term memory as meta-learners. We define global and local meta-learning variants for time series with complex seasonality and compare meta-learners on multiple forecasting problems, demonstrating their superior performance compared to simple averaging.

Deep Spectral Copula Mechanisms Modeling Coupled and Volatile Multivariate Time Series (PDF)

Yang Yang, Zhilin Zhao and Longbing Cao

Abstract:

Exploring inter- and intra-time series relations and handling volatile covariates form various challenges in modeling Coupled and Volatile Multivariate Time Series (CVMTS). A typical CVMTS data is the COVID-19 case time series across multiple countries, whose covariates may involve high volatility caused by missing samples. The existing approaches merely focus on a single set of multivariate time series or multiple multivariate time series without considering their volatile temporal covariates. They do not sufficiently characterize CVMTS features by explicitly modeling intra- and inter-MTS couplings and effectively handling volatile covariates in multiple multivariate time series. Accordingly, we propose Deep Spectral Copula Mechanisms (DSCM) to adapt CVMTS. Specifically, DSCM (1) incorporates a Singular Spectral Analysis (SSA) module to reduce the volatility of multiple covariates; (2) applies an intra-MTS coupling module to explicitly model the temporal couplings within a single set of multivariate time series; and (3) transforms target variables into joint probability distributions by Gaussian copula transformation to establish inter-MTS couplings across multiple multivariate time series. Substantial experiments on COVID-19 time-series data from multiple countries indicate the superiority of DSCM over state-of-the-art approaches.

Spatial-Temporal Residual Multi-Graph Convolution Network for Traffic Forecasting (PDF)

Ruoxuan Zhu, Yi Qian, Hui Zheng, Xing Wang, Junlan Feng, Lin Zhu and Chao Deng

Abstract:

The optimization and management of transportation play an important role in the Intelligent Transportation System (ITS), and the prediction of traffic flow data is the basis of it. Recently, graph convolution network (GCN) is widely applied to extract the features of non-Euclidean data, some models usually use the predefined distance-based adjacency matrix for GCN to extract node features. However, the distance graph can only capture part of the spatial correlation and limit the effective learning of those models. In fact, there are close similarities between the remote intersection with similar time sequence, they may be living quarters or industrial zones in the city. Besides, some previous models capture incomplete temporal correlation, which will affect the long-term prediction task to a certain extent. To address these limitations, we propose a novel Spatial-Temporal Residual Multi-Graph Convolution Network (ST-RMGCN) for traffic flow forecasting. Specifically, our model utilizes SimHash to calculate the time sequence similarity between nodes and build a semantic graph that is be instrumental in capturing remote spatial correlation. Moreover, multi-graph convolution network also extracts the features of adjacent nodes through distance graph. In addition, we design a novel time embedding method, which contains the exact time information of the time step, and it contributes to the attention mechanism and Gate Recurrent Unit (GRU) to capture the local and global time information. Extensive experiments on six real-world datasets show that our proposed model performs better than the state-of-the-art methods.

AMLNet: Adversarial Mutual Learning Neural Network for Non-AutoRegressive Multi-Horizon Time Series Forecasting (PDF)

Yang Lin

Abstract:

Multi-horizon time series forecasting, crucial across diverse domains, demands high accuracy and speed. While AutoRegressive (AR) models excel in short-term predictions, they suffer speed and error issues as the horizon extends. Non-AutoRegressive (NAR) models suit long-term predictions but struggle with interdependence, yielding unrealistic results. We introduce AMLNet, an innovative NAR model that achieves realistic forecasts through an online Knowledge Distillation (KD) approach. AMLNet harnesses the strengths of both AR and NAR models by training a deep AR decoder and a deep NAR decoder in a collaborative manner, serving as ensemble teachers that impart knowledge to a shallower NAR decoder. This knowledge transfer is facilitated through two key mechanisms: 1) outcome-driven KD, which dynamically weights the contribution of KD losses from the teacher models, enabling the shallow NAR decoder to incorporate the ensemble's diversity; and 2) hint-driven KD, which employs adversarial training to extract valuable insights from the model's hidden states for distillation. Extensive experimentation showcases AMLNet's superiority over conventional AR and NAR models, thereby presenting a promising avenue for multi-horizon time series forecasting that enhances accuracy and expedites computation.

Session: Private, Secure, and Trust Data Analytics (PSTDA III)

Defending the Graph Reconstruction Attacks for Simplicial Neural Networks (PDF)

Huixin Zhan, Liyuan Gao, Kun Zhang, Zhong Chen and Victor Sheng

Abstract:

Releasing the representations of nodes in real-world graphs associated with people or human-related activities, such as social and economic networks, gives adversaries a potential way to infer the sensitive information of edges. For example, graph convolutional layers initially aggregate node representations with their neighbors before passing them through non-linear activation functions. Hence, the released node representations may potentially breach edge privacy of the node neighbors. Thus, in this work, we study whether representations can be inverted to recover the graph used to generate them. We study three types of outputs that are trained on the graph, i.e., representations output from graph convolutional networks (GCNs), representations output from graph attention networks (GATs), and representations output from our proposed simplicial neural networks (SNNs). Unlike the first two types of representations that only encode pairwise relationships, the third type of representation, i.e., SNN outputs, encodes higher-order interactions (e.g., homological features) between nodes. We propose two graph reconstruction attacks (GRAs), i.e., Type-1 and Type-2 attacks, to recover a graph's adjacency matrix from the three types of outputs trained on the graph. Specifically, our GRAs utilize a graph-decoder to minimize the reconstruction loss for the generated adjacency matrix via backpropagation. Our conclusions are two folds. First, our Type-2 attack achieves the best performance among all current GRAs. Second, we find that GCN outputs obtain the least precision and AUC on five datasets, followed by the GAT outputs, followed by the SNN outputs. Therefore, the SNN outputs reveal the lowest privacy-preserving ability to defend the GRAs. We further propose an unbiased multi-bit rectifier, by which the server can communicate with the nodes to privately collect their representations to defend the GRAs from potential adversaries.

Underwater Localization Based on Robust Privacy-preserving and Intelligent Correction of Sound Velocity (PDF)

Jingxiang Xu, Ying Guo, Ziqi Wang, Fei Li and Ke Geng

Abstract:

The privacy-preserving localization of hydroacoustic sensor networks plays a critical role in the communication and control of marine environments. The performance of underwater location varies with constrained the complex underwater environment, such as openness, inhomogeneity, temperature, press, and so on, which make it much more challenging to ensure privacy preserving methods and obtain accurate acoustic speed used for localization computation. To address the above issues, this paper innovatively constructs Privacy-preservation Three-dimensional Underwater Location (PTUL). Firstly, the maximum distance separable coding algorithm which is designed one-time aggregated mask reconstruction by mask coding of online beacon node signals to ensure privacy-preserving and robustness is introduced into this localization model. Secondly, it relies on the sound speed modified model to compensate for the error of acoustic speed, through which an iterative regression strategy is used to deal with the change of acoustic speed. Finally, the experiments are provided to illustrate the feasibility of the proposed model. The proposed localization algorithm can efficiently improve the localization accuracy and ensure the privacy of the localization data compared with the other localization algorithms.

A Multimodal Adversarial Database: Towards a Comprehensive Assessment of Adversarial Attacks and Defenses on Medical Images (PDF)

Junyao Hu, Yimin He, Weiyu Zhang, Shuchao Pang, Ruhao Ma and Anan Du

Abstract:

Deep learning models have been widely applied in many fields, including medical image analysis and computer-aided disease diagnosis. However, these models are easily fooled by adversarial attacks from some created adversarial examples, which are hardly distinguished by humans. In this paper, we implement a comprehensive assessment of six popular adversarial attacks on four multimodal medical image datasets using two main deep learning-based target models. Moreover, to evaluate the capability of defense, two new defense methods are leveraged to cope with medical adversarial attacks. More importantly, we also build and release a big multimodal medical adversarial database (including four medical adversarial datasets) with 712,596 examples to facilitate future research of adversarial attacks and defenses in the multimodal medical image field. Extensive experiments indicate that all-sided adversarial attacks like BIM are still scarce under different evaluation metrics and defenses are not universally successful.

Enhancing Federated Learning by One-Shot Transferring of Intermediate Features from Clients (PDF)

Deng Youxingzhu, Zhou Yipeng, Liu Gang, Hui Wang and Shui Yu

Abstract:

Federated learning (FL) is an emerging paradigm using a parameter server (PS) to coordinate multiple decentralized clients for training a common model without exposing their raw data. Despite its amazing capability in preserving data privacy, FL confronts two significant challenges that have not been sufficiently addressed by existing works, which are: 1) heterogeneous data distributed on clients given that the PS cannot alter data locations. 2) limited computation resources of FL clients who may conduct model training with mobile devices. To tackle these challenges, we propose a novel federated one-shot transferring of intermediate features (FedOTF) algorithm. More specifically, FedOTF consists of two stages: feature extraction and model reconstruction. To overcome challenge 1, clients in stage 1 only collaboratively train a small model in a federated manner, with the objective to extract features. In stage 2, intermediate features (generated by the small model trained by stage 1) are transferred from clients to the PS so that a large model can be trained to overcome challenge 2. In addition, the design of FedOTF is robust, which can flexibly diminish the number of communication rounds of stage 1 when network capacity is limited, and reduce the amount of exposed features when privacy is concerned. To verify the superiority of FedOTF, we conduct comprehensive experiments with real datasets. The experiment results demonstrate that FedOTF can significantly improve the model utility of FL because a better model can be finally obtained by the PS without incurring heavy computational load on clients. Besides, we conduct robustness evaluation of FedOTF, which can achieve stable performance when varying network capacity and privacy requirement.

Session: AI and Data Science for Cybersecurity (AISC)

CRIMEO: Criminal Behavioral Patterns Mining and Extraction from Video Contents (PDF)

Raed Abdallah, Hassan Harb, Yehia Taher, Salima Benbernou and Rafiqul Haque

Abstract:

The security and well-being of a nation's citizens, as well as the protection of their lives and properties, are fundamental for prosperity. Unfortunately, in recent years, we have witnessed a surge in various types of crimes such as murder, robbery, terrorism, and kidnapping. This has placed significant pressure on Law Enforcement Agencies (LEAs) to effectively prevent and detect crimes, driving them to adopt various technologies in the criminal investigation process. Among these technologies, surveillance systems have emerged as a valuable tool for monitoring human behaviors and activities, particularly in public and densely populated areas of large cities. However, video analysis in crime investigation poses significant challenges for LEAs, requiring accurate and timely detection, recognition, and tracking of objects and individuals. Addressing this issue, we present CRIMEO, a smart system designed for mining and extracting behavioral patterns from video content. CRIMEO operates in real-time and leverages an ontology-based approach to represent complex semantic events and employ video analytics. This enables LEAs to automatically detect and identify different types of crimes. CRIMEO encompasses four key phases: data collection, analysis, storage, and visualization. In the data collection phase, video data from surveillance systems is gathered and transmitted, via Apache Kafka, to the subsequent phase after the video is split into frames. The data analysis phase applies a range of video analytics techniques, including face detection and recognition, object detection, action recognition, and more, to extract behavioral patterns from the video content. These extracted patterns are then stored in the Neo4j graph database during the data storage phase. Finally, in the visualization phase, inference rules defined by LEA experts are applied to detect and visualize criminal activities. To demonstrate the effectiveness of CRIMEO, we implemented the system across multiple scenarios, showcasing its relevance in aiding LEAs in the detection of various crime types. By utilizing CRIMEO, LEAs can benefit from advanced video analysis capabilities and real-time crime detection, ultimately enhancing their ability to maintain safety and security within their jurisdictions.

Cross-layer Federated Heterogeneous Ensemble Learning for Lightweight IoT Intrusion Detection System (PDF)

Suzan Hajj, Joseph Azar, Jacques Bou Abdo, Jacques Demerjian, Abdallah Makhoul and Dominique Ginhac

Abstract:

This paper presents a heterogeneous federated ensemble model for intrusion detection system, employing a semi-supervised novelty detection technique – the baseline K-means. The technique learns normal traffic from baseline data and utilizes the Mahalanobis distance to detect anomalous packets. To mitigate the false-positive rate inherent in anomaly-based intrusion detection system, we propose an ensemble approach that integrates local novelty detection models dedicated to each worker in both weighed and voting-based strategies. The federated design augments each worker's detection capability without increasing the false positive rate. Our extensive experiments showcase the system's robustness and adaptability over traditional standalone IDS, with marked improvements in precision, recall, and F1-score under varying sampling rates. We made this project's code publicly available on Github for replicability.

A Data-driven Approach for Risk Exposure Analysis in Enterprise Security (PDF)

Albert Calvo, Santiago Escuder, Josep Escrig, Marta Arias, Nil Ortiz and Jordi Guijarro

Abstract:

For several years, Security Operation Centers (SOCs) have relied on tools such as Security Information and Event Management (SIEM) and Intrusion Detection Systems (IDS) for reactive threat detection and risk management. However, these tools are becoming inadequate in detecting the current threat landscape, which is continuously increasing in terms of volume and variety, and targeting the most vulnerable component in the kill-chain, the human actor. This manuscript presents a novel data-driven approach that models user and entity behaviour in the early stages of the kill-chain. The proposed system estimates the probability of an entity being exposed by a threat actor during the delivery stage, thereby providing better anticipation time allowing the end-user to undertake mitigation focusing on concrete entities. Moreover, the framework has been tested in a real-life scenario executing different realistic phishing simulations and achieving successful results.

Understanding the Country-Level Security of Free Content Websites and their Hosting Infrastructure (PDF)

Mohammed Alqadhi, Ali Alkinoon, Saeed Salem and David Mohaisen

Abstract:

This paper examines free content websites (FCWs) and premium content websites (PCWs) in different countries, comparing them to general websites. The focus is on the distribution of malicious websites and their correlation with the national cyber security index (NCSI), which measures a country's cyber security maturity and its ability to deter the hosting of such malicious websites. By analyzing a dataset comprising 1,562 FCWs and PCWs, along with Alexa's top million websites dataset sample, we discovered that a majority of the investigated websites are hosted in the United States. Interestingly, the United States has a relatively low NCSI, mainly due to a lower score in privacy policy development. Similar patterns were observed for other countries With varying NCSI criteria. Furthermore, we present the distribution of various categories of FCWs and PCWs across countries. We identify the top hosting countries for each category and provide the percentage of discovered malicious websites in those countries. Ultimately, the goal of this study is to identify regional vulnerabilities in hosting FCWs and guide policy improvements at the country level to mitigate potential cyber threats.

ECC: Enhancing Smart Grid Communication with Ethereum Blockchain, Asymmetric Cryptography, and Cloud Services (PDF)

Raphaelle Akhras, Wassim El-Hajj, Hazem Hajj, Khaled Shaban and Rabih Jabr

Abstract:

Smart grids are susceptible to security vulnerabilities of cyber-physical systems due to the heterogeneity of their interconnected components. There are high risks associated with potential attacks targeting the two-way communication between the smart meters and the utility servers. It is vital to ensure that data communicated between consumers and the utility is not tampered with and is authentic, private, and available. Conventional security measures in traditional communication and network systems fail to secure the data communication aspect in the complex network that composes the advanced metering infrastructure (AMI). In this work, we propose ECC: a novel prevention approach based on Ethereum smart contracts, asymmetric cryptographic functions, and cloud services for securing the two-way communication between smart meters and utility servers. Ethereum blockchain is utilized as a building block where communicated data is treated as transactions encrypted and stored in a distributed fashion to ensure data availability, confidentiality, and privacy. We also augment the Ethereum architecture with cloud services to extend the number of allowable transactions, ensure the availability of the electricity data, and reduce the cost associated with Ethereum transactions. The conducted experiments illustrate the efficacy of ECC in terms of the achieved security properties. This paper shows that the Ethereum Blockchain coupled with Cloud services can improve the efficiency of a system solely based on the Ethereum Blockchain.

Session: Knowledge Graphs and Graph Learning (Research IV)

Knowledge Graph-based Embedding for Connecting Scholars in Academic Social Networks (PDF)

Prasad Calyam, Xiyao Cheng, Yuanxun Zhang, Harsh Joshi and Mayank Kejriwal

Abstract:

In recent years, research tasks have increasingly involved using multi-disciplinary knowledge through collaborations of scholars from multiple fields. However, identifying a team of suitable collaborators from diverse fields for a given research task is a challenging and time-consuming process. In this paper, we propose a novel "ScholarTeamFinder" model that uses knowledge graph based link prediction to identify collaborators within an academic social network (ASN) to form a research team to address a multi-disciplinary research problem. Our approach involves building a heterogeneous knowledge graph within an ASN using entities such as scholars, publications, research grants, and the relationship among these entities. Following this, we use graph-based deep learning to learn the node embedding from the knowledge graph that can be used for scholar team recommendation. More specifically, we used the classical *meth-path2vec* as our base graph learning algorithm and improved its performance by considering semantic meaning of entities and encoding edge embeddings in the graph. Finally, we propose a beam-search algorithm for scholar team prediction based on our model embeddings. Our evaluation of ScholarTeamFinder is performed using large ASN datasets including a unique dataset (i.e., NSF award dataset) of federal grant awards collected over the last ten years and the scholars' publication data, as well as three other widely used datasets (i.e., APS, SCHOLAT and Gowalla). Experiment results show that our model outperforms the state-of-the-art models across the different datasets.

Knowledge Enhanced Graph Neural Networks for Graph Completion (PDF)

Luisa Werner, Nabil Layaïda, Pierre Genevès and Sarah Chlyah

Abstract:

Graph data is omnipresent and has a wide variety of applications, such as in natural science, social networks, or the semantic web. However, while being rich in information, graphs are often noisy and incomplete. As a result, graph completion tasks, such as node classification or link prediction, have gained attention. On one hand, neural methods, such as graph neural networks, have proven to be robust tools for learning rich representations of noisy graphs. On the other hand, symbolic methods enable exact reasoning on graphs. We propose Knowledge Enhanced Graph Neural Networks (KeGNN), a neuro-symbolic framework for graph completion that combines both paradigms as it allows for the integration of prior knowledge into a graph neural network model. Essentially, KeGNN consists of a graph neural network as a base upon which knowledge enhancement layers are stacked with the goal of refining predictions with respect to prior knowledge. We instantiate KeGNN in conjunction with two well-known graph neural networks, Graph Convolutional Networks and Graph Attention Networks, and evaluate KeGNN on multiple benchmark datasets for node classification.

Lightweight Graph Convolutional Collaborative Filtering Recommendation Approach Incorporating Social Relationships (PDF)

Xiangfu Meng, Hongjin Huo, Xiaoyan Zhang and Wanchun Wang

Abstract:

Graph convolutional network (GCN) has rapidly developed in various fields due to its powerful modeling capability. However, most of the researches directly inherit the complex design of GCN, such as feature transformation and nonlinear activation, which lacks thorough ablation analysis on GCN. In addition, implicit feedback is not fully utilized and data sparsity is not well resolved, which are also shortcomings of current recommendation algorithms. To solve the above problems, this paper proposes a lightweight graph convolutional collaborative filtering (F-LightGCCF) recommendation approach incorporating social relationships. Firstly, it abandons the design of feature transformation and nonlinear activation in graph convolutional models and simplifies model training. Additionally, a series of intermediate feedback from users' implicit negative feedback is generated by taking advantage of social networks, which improves the utilization of implicit negative feedback. Secondly, it can model the long-range dependencies between users and items by using the dual attention mechanism, aggregating the contribution values of neighboring nodes and the importance of the learning vectors in each layer of the graph convolution layer respectively. Lastly, the inner product operation is used to obtain the association score between users and items. Extensive experiment results on two real-world datasets show that F-LightGCCF outperforms existing state-of-the-art recommendation methods. Further ablation studies and analyses validate the efficiency and effectiveness of the F-LightGCCF model.

Are GNNs the Right Tool to Mine the Blockchain? The Case of the Bitcoin Generator Scam (PDF)

Sam Yuen, Paula Branco, Aaron Chew, Guy-Vincent Jourdan, Fabian Lim and Laura Wynter

Abstract:

A Bitcoin Generator Scam (BGS) is a type of cyberattack in which scammers promise to provide individuals with free cryptocurrencies if they pay a mining fee. Although graph neural networks (GNNs) have been used for detecting other cryptocurrency frauds, the usefulness of these methods for BGS detection has not been studied. In this paper, we carry out extensive experiments to assess the use of both standard machine learning (ML) methods and GNNs to detect Bitcoin transactions associated with activities stemming from Bitcoin Generator Scams. We observe that the over-smoothing problem exists in GNNs designed for BGS detection and show that Random Walk Positional Encoding (RWPE) allows representing long-range interactions between far-away transactions in GNNs without causing over-smoothing. We show that the General, Powerful, Scalable (GPS) Graph Transformer with RWPE outperforms both GNN and ML based state-of-the-art fraud detection

methods in Bitcoin Generator Scams. We also analyze the effectiveness of Breadth First Search (BFS) for graph sampling and show that it should not be used as it induces bias toward the subnetwork structure. We propose the Random First Search (RFS) sampling alternative and show that this is a more suitable solution.

Session: Society and Human (Applications III)

Classification with Explanation for Human Trafficking Networks Detection (PDF)

Fabien Delorme, David Ing, Said Jabbour, Nelly Robin and Lakhdar Sais

Abstract:

On a worldwide scale, an increasing number of victims of human trafficking were observed these last years, covering a majority of countries and territories. Among them, a large portion of women and girls are recruited primarily for sexual exploitation. United Nations Office on Drugs and Crime (UNODC) highlights the difficulties of access to justice which deprive victims of protection, a central issue behind our work. Our contribution is part of an emerging research trend, combining Artificial Intelligence (AI), Humanities and Social Sciences (HSS). It makes an original use of legal database to identify Human Trafficking Networks (HTNs), involving both sexual abuse victims and exploiters. First, a reformulation of the legal database as a numerical database is proposed, using new features expressing relationships between people involved in the same court case, likely to better reveal HTNs. Secondly, six machine learning algorithms, including Decision Tree, Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) are used to train on numerical database and learn to classify the input court case into one of the three classes: Not suspicious, Suspicious, or Probably suspicious. We in details discuss knowledge-based feature engineering, dataset balancing, parameters tuning, and best models selection. The comparative empirical evaluations between those classification algorithms have been conducted to highlights the relevance of our HTNs detection approach. To help the end-users, to better understand the displayed HTNs, for Decision Tree and Random Forest, we also provide explanations of why such court case can be classified. Those results were finally discussed with experts in the field of human trafficking, providing us with interesting feedback shedding light to this multidimensional form of modern-day slavery problem.

Towards a (Semi-)Automatic Urban Planning Rule Identification in the French Language (PDF)

Maksim Koptelov, Margaux Holveck, Bruno Cremilleux, Justine Reynaud, Mathieu Roche and Maguelonne Teisseire

Abstract:

One of the objectives of the Hérelles project is to find new mechanisms to facilitate the labeling (or semantization) of clusters from time series of satellite images. To achieve this, a proposed solution is to associate textual elements of interest with satellite data. The first step in this process consists of an automatic extraction of the information in the form of rules from urban planning documents composed in the French language. To address this challenge, we propose a method which is based on the multi-label classification of textual segments. It includes a special format for representing segments, in which each segment has a title and a subtitle. In addition, we propose a cascade approach aiming to deal with hierarchy of class labels. Finally, we develop several text augmentation techniques for the texts in French, which are able to improve the prediction results. We demonstrate experimentally that the resulting framework correctly classifies each type of segment with more than 90% of accuracy.

To Personalize or Not To Personalize? Soft Personalization and the Ethics of ML for Health (PDF)

Alessandro Falcetta, Massimo Pavan, Stefano Canali, Viola Schiaffonati and Manuel Roveri

Abstract:

Personalization is among the most promising out-comes of using Machine Learning models that can be trained on data representing a specific individual. Personalization is particularly promising in areas such as health and medicine, as several crucial aspects and determinants of health are individual. Yet additional ethical issues arise with increasingly personalized models, including privacy, acceptability, reliability and trade-offs. In this paper we discuss and propose ML models for health that can be personalized on individual users, while guaranteeing both their privacy and quality from an ethical and epistemic (knowledge-related) point of view. To achieve these goals, we argue that we need to control the learning and evolution of personalized models. We propose soft personalization as an ethically-informed framework to limit personalization and respect epistemic and ethical values that are specific for the health context, including representativity, quality, non-maleficence, beneficence, privacy. Based on an interdisciplinary approach combining the philosophical and computer science scholarship of our group, soft personalization is a way of developing different models that can be selected depending on their quality and safety. We characterize the approach theoretically and technically and make it concrete with a case study of glucose monitoring and anomaly detection through privacy-preserving ML. Our framework shows that, even when individual issues such as privacy can be mitigated, trade-offs with other values remain and choices are necessary as to which values should be prioritized.

MINDSET: A benchMarking suIte exploring seNsing Data for SElf sTates inference (PDF)

Christina Karagianni, Eva Paraschou, Sofia Yfantidou and Athena Vakali

Abstract:

Ubiquitous devices, such as smartphones and wearables, are becoming increasingly popular for monitoring user behavior, health, and well-being. Through omnipresent monitoring, they aim to raise user awareness and encourage positive health behavior change. Yet, ubiquitous technologies suffer from expectation mismatch, lack of user-centric adaptiveness, and, ultimately, high abandonment rates. This work is motivated by the vital need to tailor and personalize ubiquitous technologies while dealing with the challenges arising from the lack of user profiling and the absence of relevant, self-reported user data. To this end, we show that the automatically passively collected sensing data from the wearables can be exploited to improve personalization and infer several user states relevant to demographic, physiological, psychological, and personality aspects, complementing the need for time-consuming self-reports. To accomplish this task, and enable the reproducibility and extensibility of our work; we propose an extensive benchmark suite by exploiting sensing data harvested from ubiquitous devices. Our benchmark covers a wide range of personalization tasks, including modeling gender, age, personality states, and stress, experimenting on the publicly available, newly released LifeSnaps dataset containing over 71 million rows of data capturing the daily lives of 71 participants in their naturalistic environments. The proposed benchmarking continuum showcases the strong potential for the applicability of the presented work in critical applications, such as mental healthcare monitoring, privacy preservation, and responsible artificial intelligence (AI), by fostering fairness assessments when protected attribute knowledge is unavailable.

Session: Student Competition

MAT: Effective Link Prediction via Mutual Attention Transformer (PDF)

Van Quan Nguyen, Quang Huy Pham, Quang Dan Tran, Kien Bao Thang Nguyen and Hieu Nghia Nguyen

Abstract:

The Data Science and Advanced Analytics (DSAA) 2023 competition [1] focuses on proposing link prediction methods to solve challenges about network-like data structure, such as network reconstruction, network development, etc., from articles on Wikipedia. In this challenge, our ”UIT Dark Cow” team proposes the Mutual Attention Transformer (MAT) method to predict if there is a link between two Wikipedia pages. Our method achieved the 5th and 4th position on the leaderboard for the public and private tests, respectively. Our source code is publicly available for the ease of experimental re-implementation at the following link: *https://github.com/ minhquan6203/source-code-dsaa-2023*.

Enhanced Edge Prediction, A Case Study: Predicting Links in Wikipedia Sites (PDF)

Apostolos Giannoulidis and Ioannis Mavroudopoulos

Abstract:

This study introduces a scalable approach for link prediction in Wikipedia pages, specifically designed to handle the challenges arising from the large volume of data. The proposed solution combines partial reconstruction of the original graph using node descriptions, the generation of node pair vectors based on graph metrics, and the application of a threshold similarity using the TF-IDF method. Our proposed solution achieve a high F1 score of 0.948.

Link Prediction for Wikipedia Articles as a Natural Language Inference Task (PDF)

Chau Thang Phan, Quoc-Nam Nguyen and Kiet Nguyen

Abstract:

Link prediction task is vital to automatically understanding the structure of large knowledge bases. In this paper, we present our system to solve this task at the Data Science and Advanced Analytics 2023 Competition "Efficient and Effective Link Prediction" (DSAA-2023 Competition) [1] with a corpus containing 948,233 training and 238,265 for public testing. This paper introduces an approach to link prediction in Wikipedia articles by formulating it as a natural language inference (NLI) task. Drawing inspiration from recent advancements in natural language processing and understanding, we cast link prediction as an NLI task, wherein the presence of a link between two articles is treated as a premise, and the task is to determine whether this premise holds based on the information presented in the articles. We implemented our system based on the Sentence Pair Classification for Link Prediction for the Wikipedia Articles task. Our system achieved 0.99996 Macro F1-score and 1.00000 Macro F1-score for the public and private test sets, respectively. Our team UIT-NLP ranked 3rd in performance on the private test set, equal to the scores of the first and second places. Our code1 is publicly for research purposes.

A Text-based Approach For Link Prediction on Wikipedia Articles (PDF)

Anh Tran, Tam Nguyen and Son Luu

Abstract:

This paper present our work in the DSAA 2023 Challenge about Link Prediction for Wikipedia Articles. We use traditional machine learning models with POS tags (part-of-speech tags) features extracted from text to train the classification model for predicting whether two nodes has the link. Then, we use these tags to test on various machine learning models. We obtained the results by F1 score at 0.99999 and got 7th place in the competition. Our source code is publicly available at this link: https://github.com/Tam1032/ DSAA2023-Challenge-Link-prediction-DS-UIT SAT.

Link Prediction on Graphs Using NLP Embedding (PDF)

João Victor Galvão da Mata and Martin Skovgaard Andersen

Abstract:

This paper summarizes our approach (team MathLand) to the DSAA 2023 Data Science Competition, which focuses on link prediction. Our proposed model is based on embedding techniques commonly used for natural language processing, and the embedding is constructed as part of the neural network training, eliminating the need for a separate step. We train the model using the binary cross entropy loss and the Adam optimizer. The approach achieves high F1-scores on validation and test sets.

Predict Link Between Nodes Using An Ensemble Modelling Combining Depth Search Algorithm And Textual Similarity Score (PDF)

Aditya Kansal and Rishabh Mehta

Abstract:

Link prediction is an important task applied in networks. Given a pair of nodes (u, v) we would need to predict if the edge between nodes u and v will be present or not. Link prediction is strongly related to recommendations, network reconstruction and network evolution. In this paper, we focus on the link prediction task applied to Wikipedia articles. Given a sparsed subgraph of the Wikipedia network, we need to predict if a link exists between two Wikipedia pages u and v.

Achieving High Performance in Link Prediction for Wikipedia Articles Using Ensemble Approach (PDF)

Weiwu Yang

Abstract:

Link prediction is a crucial task in graph mining, with numerous applications in social networks, product and movie recommendation and knowledge graph completion. While recent research has focused on Graph Neural Network (GNN) architectures, traditional machine learning methods can also achieve comparable performance in certain cases. In this paper, we present our solution that ranked sixth in the “Link Prediction for Wikipedia Articles” competition at DSAA2023. We used three distinct approaches for link prediction: a tree-based gradient boosting method, a tree-based ensemble learning method, and a neural network method. We ensembled these methods using a voting scheme to determine the final prediction results. Our model achieved F1-scores of over 0.99 on most prediction results.

Session: Feature and Label Learning (Research V)

Sample Topology Exploration for Label Distribution Learning (PDF)

Yan-Wen Xiong, Heng-Ru Zhang, Fan Min and Peng-Cheng Li

Abstract:
Label distribution learning (LDL) employs probabilistic labels to capture the varying degrees of relevance among decision attributes. Existing LDL algorithms usually employ sample correlations to improve their predictive validity. However, they merely employ the superficial features of the sample for correlation analysis, seldom delving into its latent deep features. In this paper, we propose an algorithm to explore the sample topology (ST-LDL) to address this problem. First, we construct a locally weighted directed graph for each target sample. The target sample and its k neighbors are regarded as nodes within the graph. The asymmetrical correlation among these k+1 samples is computed separately to determine the weight value of the graph. Then, we utilise the local topology between samples to mine latent information and reconstruct the samples. The local graph is fed into the graph convolutional network to nonlinearly reconstruct the features by mining the latent information. Finally, we designed a new optimization objective function for the reconstructed samples. Experiments are carried out on elven real-world datasets in comparison with seven state-of-the-art algorithms. The results show that our algorithm outperforms several other algorithms, proving the effectiveness of our algorithm.

Causal Feature Selection: Methods and a Novel Causal Metric Evaluation Framework (PDF)

Rezaur Rashid, Jawad Chowdhury and Gabriel Terejanu

Abstract:
The proliferation of high-dimensional data in the era of big data has presented significant challenges for machine learning models. Feature selection methods have emerged as essential preprocessing techniques to address these challenges. However, most existing feature selection techniques primarily rely on correlations or associations between features and the target variable, overlooking the consideration of causal relationships. This study introduces a novel causal feature selection (CFS) algorithm that leverages causal structure learning to identify a subset of causal features. Our approach involves employing a causal graph discovery method to represent the causal relationships among variables and the causal effects of features on the target variable. To evaluate the effectiveness of our proposed CFS algorithm, we introduce a new evaluation criterion based on causal metrics, offering a principled and rigorous approach to assess the performance of causal feature selection methods. We empirically evaluate our algorithm using synthetic and real-world datasets, demonstrating that the truncated subsets of features selected by the CFS algorithm exhibit comparable or improved performance compared to baseline methods while utilizing fewer causal features.

ProPML: Probability Partial Multi-label Learning (PDF)

Łukasz Struski, Adam Pardyl, Jacek Tabor and Bartosz Zieliński

Abstract:
Partial Multi-label Learning (PML) is a type of weakly supervised learning where each training instance corresponds to a set of candidate labels, among which only some are true. In this paper, we introduce ProPML, a novel probabilistic approach to this problem that extends the binary cross entropy to the PML setup. In contrast to existing methods, it does not require suboptimal disambiguation and, as such, can be applied to any deep architecture. Furthermore, experiments conducted on artificial and real-world datasets indicate that ProPML outperforms existing approaches, especially for high noise in a candidate set.

CaFe DBSCAN: A Density-based Clustering Algorithm for Causal Feature Learning (PDF)

Pascal Weber, Lukas Miklautz, Akshey Kumar, Claudia Plant and Moritz Grosse-Wentrup

Abstract:
Causal Feature Learning (CFL) infers macro-level causes (e.g., an aggregation of pixels in a traffic light image) from micro-level data (e.g., pixels of the image) by clustering the predicted probabilities of effect states (e.g., state of the traffic light). The current method for CFL uses a two-step procedure. First, a classifier for the effect states is trained, and afterwards, the predicted effect state probabilities are clustered. With CAFE DBSCAN, we present a novel density-based clustering method that conducts CFL directly by estimating conditional probabilities during clustering. To this end, we introduce the notion of clustering regions with similar

conditional probabilities of the effect states given their micro-level data points. Our single-step approach has the following benefits: (1) CAFE DBSCAN introduces a comprehensive approach to Causal Feature Learning. Unlike existing methods, CAFE DBSCAN uses a probabilistic framework and does not require separate classification and clustering steps implemented by different algorithms relying on various assumptions, parameter settings, and optimization goals. (2) We do not need to train and tune a classifier first, hence the algorithm is more runtime-efficient than the current approach. (3) Due to the properties of density-based clustering algorithms, CAFE DBSCAN is robust against noise and outliers, which leads to purer clusters. (4) Our algorithm automatically infers a reasonable number of clusters, i.e., macro-level causes. We demonstrate the benefits of CAFE DBSCAN on synthetic and real-world data.

Session: Science and Environment (Applications IV)

Disaster Image Classification Using Pre-trained Transformer and Contrastive Learning Models (PDF)

Soudabeh Taghian Dinani and Doina Caragea

Abstract:

Natural disasters can have devastating consequences for communities, causing loss of life and significant economic damage. To mitigate these impacts, it is crucial to quickly and accurately identify situational awareness and actionable information useful for disaster relief and response organizations. In this paper, we study the use of advanced transformer and contrastive learning models for disaster image classification in a humanitarian context, with focus on state-of-the-art pre-trained vision transformers such as ViT, CSWin and a state-of-the-art pre-trained contrastive learning model, CLIP. We evaluate the performance of these models across various disaster scenarios, including in-domain and cross-domain settings, as well as few-shot learning and zero-shot learning settings. Our results show that the CLIP model outperforms the two transformer models (ViT and CSWin) and also ConvNeXts, a competitive CNN-based model resembling transformers, in all the settings. By improving the performance of disaster image classification, our work can contribute to the goal of reducing the number of deaths and economic losses caused by disasters, as well as helping to decrease the number of people affected by these events.

Non-Redundant Image Clustering of Early Medieval Glass Beads (PDF)

Lukas Miklautz, Andrii Shkabrii, Collin Leiber, Bendeguz Tobias, Benedict Seidl, Elisabeth Weissensteiner, Andreas Rausch, Christian Böhm and Claudia Plant

Abstract:

Glass beads were among the most common grave goods in the Early Middle Ages, with an estimated number in the millions. The color, size, shape and decoration of the beads are diverse leading to many different archaeological classification systems that depend on the subjective decisions of individual experts. The lack of an agreed upon expert categorization leads to a pressing problem in archaeology, as the categorization of archaeological artifacts, like glass beads, is important to learn about cultural trends, manufacturing processes or economic relationships (e.g., trade routes) of historical times. An automated, objective and reproducible classification system is therefore highly desirable. We present a high-quality data set of images of Early Medieval beads and propose a clustering pipeline to learn a classification system in a data-driven way. The pipeline consists of a novel extension of deep embedded non-redundant clustering to identify multiple, meaningful clustering of glass bead images. During the cluster analysis we address several challenges associated with the data and as a result identify high-quality clustering that overlap with archaeological domain expertise. To the best of our knowledge this is the first application of non-redundant image clustering for archaeological data.

Exploring Deep Learning for Full-disk Solar Flare Prediction with Empirical Insights from Guided Grad-CAM Explanations (PDF)

Chetraj Pandey, Anli Ji, Trisha Nandakumar, Rafal Angryk and Berkay Aydin

Abstract:

This study progresses solar flare prediction research by presenting a full-disk deep-learning model to forecast ≥M class solar flares and evaluating its efficacy on both central (within ±70◦) and near-limb (beyond ±70◦) events, showcasing qualitative assessment of post hoc explanations for the model's predictions, and providing empirical findings from human-centered quantitative assessments of these explanations. Our model is trained using hourly full-disk line-of-sight magnetogram images to predict ≥M-class solar flares within the subsequent

24-hour prediction window. Additionally, we apply the Guided Gradient-weighted Class Activation Mapping (Guided Grad-CAM) attribution method to interpret our model’s predictions and evaluate the explanations. Our analysis unveils that full-disk solar flare predictions correspond with active region characteristics. The following points represent the most important findings of our study: (1) Our deep learning models achieved an average true skill statistic (TSS) of ∼0.51 and a Heidke skill score (HSS) of ∼0.38, exhibiting skill to predict solar flares where for central locations the average recall is ∼0.75 (recall values for X- and M-class are 0.95 and 0.73 respectively) and for the near-limb flares the average recall is ∼0.52 (recall values for X- and M- class are 0.74 and 0.50 respectively); (2) qualitative examination of the model’s explanations reveals that it discerns and leverages features linked to active regions in both central and near-limb locations within full-disk magnetograms to produce respective predictions. In essence, our models grasp the shape and texture-based properties of flaring active regions, even in proximity to limb areas—a novel and essential capability with considerable significance for operational forecasting systems.

Utilizing MODIS Fire Mask for Predicting Forest Fires Using Landsat-9/8 and Meteorological Data (PDF)

Yash Gupta, Navneet Goyal, Vishal John Varghese and Poonam Goyal

Abstract:

Recent years have seen some of the largest forest fires ever, including the 2020 California megafires and the Australian bushfires, causing billions of dollars in property damage and destroying millions of acres of green reserves. The subject of forest fires becomes even more alarming when viewed in conjunction with the increasingly concerning problems of climate change and global warming. The planning regarding prevention and mitigation of forest fires and management of nearby areas can greatly benefit from an accurate prediction model. The objective of this study is to develop deep learning models, which use satellite images and meteorological data to pinpoint potential fires at a pixel granularity. Data from the recently launched Landsat-8 and Landsat-9 satellite systems have been used to predict forest fires at a spatial resolution of 30m.

The proposed solution uses the comprehensive geographical, meteorological, and MODIS-based fire history of the region, integrated from different data sources with pixel-level reprojection, as a multivariate time series (MVTS) to model the prediction problem as a binary classification problem. We adopt an encoder-classifier architecture: the BiLSTM-attention-based encoder is trained with supervised contrastive learning, while the fully-connected classifier is optimized against a weighted loss for increased recall. Our experiments demonstrate that the proposed model is robust to spatial and temporal variations in occurrence of fires, thereby making its deployment possible in any region of the world. With a mean AUC of 0.99, our proposed model outperforms the existing forest fire prediction models.

Session: Practical Applications of Explainable Artificial Intelligence Methods (PRAXAI I)

Towards Explaining Satellite Based Poverty Predictions with Convolutional Neural Networks (PDF)

Hamid Sarmadi, Thorsteinn Rögnvaldsson, Mattias Ohlsson, Nils Roger Carlsson, Ibrahim Wahab and Ola Hall

Abstract:

Deep convolutional neural networks (CNNs) have been shown to predict poverty and development indicators from satellite images with surprising accuracy. This paper presents a first attempt at analyzing the CNNs responses in detail and explaining the basis for the predictions. The CNN model, while trained on relatively low resolution day- and night-time satellite images, is able to outperform human subjects who look at high-resolution images in ranking the Wealth Index categories. Multiple explainability experiments performed on the model indicate the importance of the sizes of the objects, pixel colors in the image, and provide a visualization of the importance of different structures in input images. A visualization is also provided of type images that maximize the network prediction of Wealth Index, which provides clues on what the CNN prediction is based on.

Text Classification is Keyphrase Explainable! Exploring Local interpretability of Transformer Models with Keyphrase Extraction (PDF)

Dimitrios Akrivousis, Nikolaos Mylonas, Ioannis Mollas and Grigorios Tsoumakas

Abstract:

Keyphrase extraction is a widely discussed topic in Natural Language Processing, as it offers a concise summary of the main topics in a document. Interpretability is also an important aspect in Machine Learning as it helps prevent socio-ethical issues, such as bias and discrimination against minorities, or mistakes that may have serious consequences. Interpretability has recently gained prominence in the field of Natural Language Processing, where transformers are the dominant architectures. The goal of interpretability is to provide interpretations that pinpoint the elements of an instance contributing the most to its decision. In this work, we use keyphrase extraction to facilitate the interpretability process, producing smaller, more concise interpretations that also consider word interactions, as keyphrases usually consist of multiple words. Additionally, our technique is based on semantic similarity, making it faster and zero-shot ready, which is ideal for online learning scenarios. We evaluated the effectiveness of our technique through a series of quantitative and qualitative experiments on the well-known BERT model, comparing it against several state-of-the-art competitors.

Interpreting Black-box Machine Learning Models for High Dimensional Datasets (PDF)

Md. Rezaul Karim, Md. Shajalal, Alex Graß, Till Döhmen, Sisay Adugna Chala, Christian Beecks and Stefan Decker

Abstract:

Many datasets are of increasingly high dimensionality, where a large number of features could be irrelevant to the learning task. The inclusion of such features would not only introduce unwanted noise but also increase computational complexity. Deep neural networks (DNNs) outperform machine learning (ML) algorithms in a variety of applications due to their effectiveness in modelling complex problems and handling high-dimensional datasets. However, due to non-linearity and higher-order feature interactions, DNN models are unavoidably opaque, making them *black-box* methods. In contrast, an interpretable model can identify statistically significant features and explain the way they affect the model's outcome. In this paper, we propose a novel method to improve the interpretability of black-box models in the case of high-dimensional datasets. First, a black-box model is trained on full feature space that learns useful embeddings on which the classification is performed. To decompose the inner principles of the black-box and to identify top-k important features (global explainability), probing and perturbing techniques are applied. An interpretable *surrogate model* is then trained on top-k feature space to approximate the black-box. Finally, decision rules and counterfactuals are derived from the surrogate to provide local decisions. Our approach outperforms tabular learners, e.g., TabNet and XGboost, and SHAP-based interpretability techniques, when tested on a number of datasets having dimensionality between 54 and 20,531.

Enhanced Explanations for Knowledge-Augmented Clustering Using Subgroup Discovery (PDF)

Maciej Szelążek, Daniel Hudson, Szymon Bobek, Grzegorz J. Nalepa and Martin Atzmueller

Abstract:

Contemporary machine learning techniques are capable of extracting complex structure from data in a way that complements or exceeds manual examination, yet, as is well– documented, many of these techniques suffer from a lack of interpretability. This paper extends previous work on explainable and interpretable machine learning, in particular on the 'Knowledge–Augmented Clusters (KnAC)' approach, allowing human users to benefit from uninterpretable 'black box' models to extract structure from datasets by clustering and to make this better understandable. One of the key functions of KnAC is to relate expert–annotated clusters to clusters that have been identified by a machine learning method, and then provide a comprehensible explanation, thus clarifying the relationships that KnAC discovered. Our novel contribution in this paper is to examine the usefulness of subgroup discovery as a way to generate comprehensible explanations within KnAC, and to compare this to the existing approach based on the XAI algorithm Anchors through a detailed evaluation. We find that the approach using subgroup discovery performs equally or better in our extensive experimentation testing this on six different datasets.

Session: Medicine (Applications V)

A Framework for Context-Sensitive Prediction in Time Series – Feasibility Study for Data-Driven Simulation in Medicine (PDF)

Fatoumata Dama, Christine Sinoquet and Corinne Lejus-Bourdeau

Abstract:

The need to comprehend and predict the dynamics of complex systems has spurred developments of time series forecasting methods across several disciplines. Nowadays, time series are collected with event logs for an ever increasing number of systems. Event logs can contain prominent information about the system dynamics. This paper addresses joint modelling of time series and event traces, to enhance time series forecasting. We introduce the Non-Homogeneous Markov Chain AutoRegressive model, to best apprehend the combined influence of past events belonging to several event categories, on a system's dynamics. The originality of our proposal stems from the synchronization of a Hawkes temporal point process with the classical first-order hidden Markov model, through contextual variables. We also instantiate a basic version whose contextual variables only take into account events' latest occurrences. Our proof-of-concept experiments address a real-world case related to digitally assisted training in anaesthesiology. We demonstrate that the advanced instantiation outperforms the basic instantiation (maximal prediction error percentage: 5.6% versus 13.5%). Finally, we validate the suitability of the advanced instantiation for the desired simulation, by applying real sequences of medical actions to digital patients. We show that we obtain highly realistic simulations.

Optimizing Resource Allocation for Tumor Simulations over HPC Infrastructures (PDF)

Errikos Streviniotis, Nikos Giatrakos, Yannis Kotidis, Thalia Ntiniakou and Miguel Ponce de Leon

Abstract:

We introduce RATS (Resource Allocator for Tumor Simulations), the first optimizer for the execution of tumor simulations over HPC infrastructures. The optimization framework of RATS incorporates 3 vital performance criteria (i) expected utility of a simulation in terms of effective drug combination on the simulated tumor, (ii) simulation execution time and (iii) number of cores required for achieving that execution time. RATS is to be used by life scientists at the Barcelona Supercomputing Center to not only remove the burden of blindly guessing the core hours we need to reserve from HPC admins to study various tumor treatment methodologies, but also to help in more rapidly distinguishing effective drug combinations, thus, potentially cutting time to market for new cancer therapies.

Death after Liver Transplantation: Mining Interpretable Risk Factors for Survival Prediction (PDF)

Veronica Guidetti, Giovanni Dolci, Erica Franceschini, Erica Bacca, Giulia Burastero, Davide Ferrari, Valentina Serra, Fabrizio Di Benedetto, Cristina Mussini and Federica Mandreoli

Abstract:

This study introduces a novel approach to mine risk factors for short-term death after liver transplantation (LT). The method outputs intelligible survival models by combining Cox's regression with a genetic programming technique known as multi-objective symbolic regression (MOSR). We consider 485 Electronic Health Records (EHRs) of patients who underwent LT, containing information on hospitalization and preoperative conditions, with a focus on infections and colonizations by multi-resistant Gram-negative bacteria. We evaluate MOSR outcomes against several performance metrics and demonstrate that they are well-calibrated, predictive, safe, and parsimonious. Finally, we select the most promising post-LT early survival risk score based on information criteria, performance, and out-of-distribution safety. Validating this technique at a multi-center level could improve service pipeline logistics through a trustworthy machine-learning method.

Session: Journal I

TOCOL: Improving Contextual Representation of Pre-trained Language Models via Token-Level Contrastive Learning (PDF)

Keheng Wang, Chuantao Yin, Rumei Li, Sirui Wang, Yunsen Xian, Wenge Rong and Zhang Xiong

Abstract:

Self-attention, which allows transformers to capture deep bidirectional contexts, plays a vital role in BERT-like pre-trained language models. However, the maximum likelihood pre-training objective of BERT may produce an anisotropic word embedding space, which leads to biased attention scores for high-frequency tokens, as they are very close to each other in representation space and thus have higher similarities. This bias may ultimately affect the encoding of global contextual information. To address this issue, we propose TOCOL, a TOken-Level COntrastive Learning framework for improving the contextual representation of pre-trained language models, which integrates a novel self-supervised objective to the attention mechanism to reshape the word representation space and encourages PLM to capture the global semantics of sentences. Results on the GLUE Benchmark show that TOCOL brings considerable improvement over the original BERT. Furthermore, we conduct a detailed analysis and demonstrate the robustness of our approach for low-resource scenarios.

GS2P: A Generative Pre-trained Learning to Rank Model with Over-parameterization for Web-Scale Search (PDF)

Yuchen Li, Haoyi Xiong, Linghe Kong, Jiang Bian, Shuaiqiang Wang, Guihai Chen and Dawei Yin

Abstract:

While *learning to rank* (LTR) is widely employed in web searches to prioritize pertinent webpages from the retrieved contents based on input queries, traditional LTR models stumble over two principal stumbling blocks leading to subpar performance: 1) the lack of well-annotated query-webpage pairs with ranking scores to cover search queries of various popularity, debilitating their coverage of search queries across the popularity spectrum, and 2) ill-trained models that are incapable of inducing generalized representations for LTR, culminating in overfitting. To tackle the above challenges, we proposed a *Generative Semi-Supervised Pre-trained* (GS2P) Learning to Rank model. Specifically, GS2P first generates pseudo-labels for the unlabeled samples using tree-based LTR models after a series of co-training procedures, then learns the representations of query-webpage pairs with self-attentive transformers via both discriminative (LTR) and generative (denoising autoencoding for reconstruction) losses. Finally, GS2P boosts the performance of LTR through incorporating Random Fourier Features to over-parameterize the models into "interpolating regime", so as to enjoy the further descent of generalization errors with learned representations. We conduct extensive offline experiments on a publicly available dataset and a real-world dataset collected from a large-scale search engine. The results show that GS2P can achieve the best performance on both datasets, compared to baselines. We also deploy GS2P at a large-scale web search engine with realistic traffic, where we can still observe significant improvement in real-world applications. GS2P performs consistently in both online and offline experiments.

PANACEA: A Neural Model Ensemble for Cyber-Threat Detection (PDF)

Malik Al-Essa, Giuseppina Andresini, Annalisa Appice and Donato Malerba

Abstract:

This study describes a new cyber-threat detection method, named PANACEA, that uses Ensemble Deep Learning coupled with Adversarial Training and XAI, to gain accuracy with neural models trained in cybersecurity problems.

Session: Practical Applications of Explainable Artificial Intelligence Methods (PRAXAI II)

Towards Quality Measures for xAI algorithms: Explanation Stability (PDF)

Marek Pawlicki

Abstract:

The domain of Artificial Intelligence has become ubiquitous across a wide plethora of domains and is now an integral part of the daily life of the ordinary citizen. While the need for increased transparency of the highly accurate black-box model is an important and very active area of research, the produced explanations themselves might not always be accurate. The measures to assess the quality of explanations are an important research topic. In this paper, a set of extensive experiments is performed to evaluate the stability of SHAP explanations under conditions of different noise types and different noise intensities as an effort to build a formal way of assessing the quality of explanations provided by the SHAP algorithm. The experiments are performed on four different datasets, with three different noise types at four different strength levels. The impact of the scenarios on SHAP explanations is reported, the implications for the evaluation of explainability methods are elaborated upon, along with the significance of the results for the SHAP method of explanations. The future directions are laid out thereafter.

ORANGE: Opposite label soRting for tANGent Explanations in heterogeneous spaces (PDF)

Alejandro Kuratomi, Zed Lee, Ioanna Miliou, Tony Lindgren and Panagiotis Papapetrou

Abstract:

Most real-world datasets have a heterogeneous feature space composed of binary, categorical, ordinal, and continuous features. However, the currently available local surrogate explainability algorithms do not consider this aspect, generating infeasible neighborhood centers which may provide erroneous explanations. To overcome this issue, we propose ORANGE, a local surrogate explainability algorithm that generates high-

accuracy and high-fidelity explanations in heterogeneous spaces. ORANGE has three main components: (1) it searches for the closest feasible counterfactual point to a given instance of interest by considering feasible values in the features to ensure that the explanation is built around the closest feasible instance and not any, potentially non-existent instance in space; (2) it generates a set of neighboring points around this close feasible point based on the correlations among features to ensure that the relationship among features is preserved inside the neighborhood; and (3) the generated instances are weighted, firstly based on their distance to the decision boundary, and secondly based on the disagreement between the predicted labels of the global model and a surrogate model trained on the neighborhood. Our extensive experiments on synthetic and public datasets show that the performance achieved by ORANGE is best-in-class in both explanation accuracy and fidelity.

Instils Trust in Random Forest Predictions (PDF)

Gopal Jamnal and Gopal Jamnal

Abstract:

This paper addresses the interpretability and transparency behind the random forest model predictions. Random forest is an ensemble of bootstrapped independent decision trees that are trained on subsets of input data to make predictions. Although random forest is a robust model that can overcome bias, its inherent complexity, and poor interpretability can make it challenging to apply in many application domains that require transparency and explainability in the model's predictions. This lack of transparency in the decision-making process can prevent users from analyzing what makes the model arrive at a specific prediction. The paper presents a visual analytic application to overcome transparency challenges by providing a clear structure of individual decision trees and hierarchical relationships between features. This allows users to analyze latent information and instils trust in the random forest model. Additionally, statistical analysis of feature ranking agreement and prediction popularity reduces mental burden of the user. The paper includes two case studies to evaluate model's uncertainty, bias, and variances in predictions with explainability at local and global scales of decision paths. The visual analytics application provides a coordinated multiple-view system to instil trust in random forest models.

Session: Learning Methods and Theories (Research VI)

Adaptive Clustered Federated Learning with Representation Similarity (PDF)

Chiyu Cai, Wei Wang and Yuan Jiang

Abstract:

Federated learning is a promising machine learning paradigm that enables participating clients to train models collaboratively with privacy restrictions. However, one of the most challenging problems in federated learning is that local data on the clients might come from different distributions. Such data heterogeneity among the clients might influence the performance of federated learning methods. In this paper, we propose FedACRS, an algorithm that deals with heterogeneous data by clustering clients with similar data distributions into groups and then performing federated learning within each group. FedACRS measures the similarity between the clients in every round based on the representation similarity and then adaptively discovers the clustering structure among the clients. To cluster the clients appropriately, we provide theoretical analysis to help determine the number of potential clusters. The results of extensive experiments in different settings demonstrate the advantage of FedACRS over the compared methods.

Learning Representations through Contrastive Strategies for a more Robust Stance Detection (PDF)

Udhaya Kumar Rajendran, Amine Trabelsi and Amir Ben Khalifa

Abstract:

Stance Detection refers to the process of determining an author's position towards a particular issue or target in a text. Previous research suggests that existing systems for Stance Detection are not resilient enough to handle variations and errors in input sentences. In our proposed methodology, we utilize Contrastive Learning to learn sentence representations. We achieve this by bringing semantically similar sentences and those implying the same stance closer to each other in the embedding space. To compare our approach, we use a pre-trained transformer model that is directly fine-tuned with the stance datasets. We evaluate the resilience of the models using char-level and word-level adversarial perturbation attacks and show that our approach performs better and is more robust to the different adversarial perturbations introduced to the test data. Our approach is also shown to perform better on small-sized and class-imbalanced stance datasets. We further experiment with unlabeled stance datasets to make the representation learning independent of domain-specific labels, and the models trained with our approach on unlabeled datasets are still robust and perform comparably to those trained with labeled data.

Toward a Realistic Benchmark for Out-of-Distribution Detection (PDF)

Pietro Recalcati, Fabio Garcea, Luca Piano, Fabrizio Lamberti and Lia Morra

Abstract:

Deep neural networks are increasingly used in a wide range of technologies and services, but remain highly susceptible to out-of-distribution (OOD) samples, that is, drawn from a different distribution than the original training set. A common approach to address this issue is to endow deep neural networks with the ability to detect OOD samples. Several benchmarks have been proposed to design and validate OOD detection techniques. However, many of them are based on far-OOD samples drawn from very different distributions, and thus lack the complexity needed to capture the nuances of real-world scenarios. In this work, we introduce a comprehensive benchmark for OOD detection, based on ImageNet and Places365, that assigns individual classes as in-distribution or out-of-distribution depending on the semantic similarity with the training set. Several techniques can be used to determine which classes should be considered in-distribution, yielding benchmarks with varying properties. Experimental results on different OOD detection techniques show how their measured efficacy depends on the selected benchmark and how confidence-based techniques may outperform classifier-based ones on near-OOD samples.

On the Independence of Adversarial Transferability to Topological Changes (PDF)

Carina Newen and Emmanuel Müller

Abstract:

One curious property of neural networks is the vulnerability to specific attacks, often called adversarial examples. One of the directions adversarial transferability research has taken is to focus on dataset features. The transferability of adversaries is often linked to those common global features being present or not. To validate this theory, we tested if the transferability of attacks occurs when the underlying global features of a dataset remain the same. This is because topology promises to preserve the properties of an object under continuous deformations. In this paper, we test the correlation between topological similarities using the mapper algorithm by Singh et al. to generate an approximation of the topology in a graphical manner and a distance notion provided by the NetLSD algorithm, which promises size, scale, and permutation invariance. These two algorithms allow us to show that adversarial transferability is, in fact, independent of the topological similarity of datasets. We implement our findings in https://github.com/KDD-OpenSource/Topological Transf. This is an astounding new insight, as former theories have led us to expect that if the assumption is true that global features are relevant for transferability, those should be captured using algorithms that detect global features under only topological change- Unless, of course, the transferability and those global features are explicitly agnostic to topological change. This might point to current research regarding adversarial transferability in different directions. More specifically, we take an experimental approach using topological approximation methods to capture essential features of datasets. Past studies concerning adversarial examples show that attacks can transfer in unforeseen ways and between different neural network architectures and may produce severe vulnerabilities in sophisticated learners. However, when tackling the problem of vulnerabilities to adversarial attacks, only a few approaches find generalizable results, and by no means have we answered when and how to attack transferability can occur. This paper shows that if we limit changes in a dataset to topological permutations, the transferability of adversarial examples generated will stay the same regardless of the amount of topological change. Since acceptance of the paper, we have actually extended our implementation to other adversarial methods by simply including given code from more methods into the general implementation. The code base is also easily extendable to other datasets for further reproducibility.

Session: Industrial

Prioritization of Identified Data Science Use Cases in Industrial Manufacturing via C-EDIF Scoring (PDF)

Raphael Fischer, Andreas Pauly, Rahel Wilking, Anoop Kini and David Graurock

Abstract:

While data science and artificial intelligence (AI) can be highly beneficial for industrial manufacturers, it is not yet readily usable. Therefore, putting it to good use requires to understand the domain challenges and identify opportunities for deploying AI. Our work aims at solving this task by proposing a generalized framework for (1) exploring companies for use cases and (2) prioritizing them via C-EDIF scoring. This novel approach allows to determine the business importance of any use case by considering the underlying evaluability, data situation, impact and infeasibility. Besides the theoretical framework, our work also provides real-world insights from applying C-EDIF scoring in an extensive use case exploration phase. These results stem from a strategic partnership between data scientists and Wilo SE, a renowned pump manufacturing company, where we successfully identified and rated opportunities for AI.

Opportunistic Air Quality Monitoring and Forecasting with Expandable Graph Neural Networks (PDF)

Jingwei Zuo, Wenbin Li, Michele Baldo and Hakim Hacid

Abstract:

Air Quality Monitoring and Forecasting has been a popular research topic in recent years. Recently, data-driven approaches for air quality forecasting have garnered significant attention, owing to the availability of well-established data collection facilities in urban areas. Fixed infrastructures, typically deployed by national institutes or tech giants, often fall short in meeting the requirements of diverse personalized scenarios, e.g., forecasting in areas without any existing infrastructure. Consequently, smaller institutes or companies with limited budgets are compelled to seek tailored solutions by introducing more flexible infrastructures for data collection. In this paper, we propose an expandable graph attention network (EGAT) model, which digests data collected from existing and newly-added infrastructures, with different spatial structures. Additionally, our proposal can be embedded into any air quality forecasting models, to apply to the scenarios with evolving spatial structures. The proposal is validated over real air quality data from PurpleAir.

Short-term Forecast and Long-term Simulation for Accurate Energy Consumption Prediction (PDF)

Daniele Giampaoli, Francesca Cipollini, Denise Maffione and Luca Oneto

Abstract:

Accurate energy consumption forecasting has become pivotal for many companies as a way to tailor the budget dedicated to energy purchase on their actual power demand, thus sustainably minimizing energy waste and expenses. For these companies, both short-term and long-term energy consumption forecasts are a matter of interest since they would like to both program last-minute buy and sell and also plan future investments for power optimization. For this purpose, in this paper, different Deep Neural Networks techniques will be tested to perform both a supervised short-term energy consumption forecasting and an unsupervised long-term simulation via generative learning since very long-term forecasting (i.e., more than 1 year) is usually too inaccurate. The first task will be performed by adopting both a Recurrent Neural Network and a Long Short-Term Memory Network, while the second one will be performed by adopting a Generative Adversarial Network. Result on public data from the Australian Energy Market Operator will support the proposal.

Practical Insights on Incremental Learning of New Human Physical Activity on the Edge (PDF)

Georgios Arvanitakis, Jingwei Zuo, Mthandazo Ndhlovu and Hakim Hacid

Abstract:

Edge Machine Learning (Edge ML), which shifts computational intelligence from cloud-based systems to edge devices, is attracting significant interest due to its evident benefits including reduced latency, enhanced data privacy, and decreased connectivity reliance. While these advantages are compelling, they introduce unique challenges absent in traditional cloud-based approaches. In this paper, we delve into the intricacies of Edge-based learning, examining the interdependencies among: (i) constrained data storage on Edge devices, (ii) limited computational power for training, and (iii) the number of learning classes. Through experiments conducted using our MAGNETO system, that focused on learning human activities via data collected from mobile sensors, we highlight these challenges and offer valuable perspectives on Edge ML.

Session: Journal II

Hyperparameter Analysis of Wide-Kernel CNN Architectures in Industrial Fault Detection – An Exploratory Study (PDF)

Jurgen van den Hoogen, Dan Hudson, Stefan Bloemheuvel and Martin Atzmueller

Abstract:

In recent years, industrial fault detection has become more data-driven due to advancements in automated data analysis using Deep Learning (DL). These techniques facilitate meaningful feature extraction, e.g., in time series data retrieved from sensors, which is typically of complex nature. This enables effective fault detection and prognostics, which increases efficiency and productivity of industrial equipment. However, the optimal settings for these DL architectures are generally use-case specific. This paper is an extended abstract of our work [1] that explores the influence of various architectural hyperparameters on the performance of one-dimensional convolutional neural networks (CNN). Using a multi-method approach, this paper focuses specifically on wide-kernel (WK) CNN models for industrial fault detection, that have proven to perform well for these tasks. By altering hyperparameters of the model's architecture such as the kernel size, stride and number of filters, an extensive hyperparameter space search was conducted. In total, we tested 12,960 different combinations for three different datasets, all representing vibration data from industrial equipment, into a model hyperparameter dataset, along with their respective performance on the underlying fault detection task. We explore these different hyperparameter settings to determine which configurations generate a good or bad result, and to obtain insights resembling 'general rules' for applying a one-dimensional WK-CNN in a signal classification context. In particular, we employ the model architecture previously proposed by [2,3], which is based on the wide-kernel framework developed by [4]. This architecture has proven to achieve state-of-the-art performance in fault detection tasks [2–5]. Also, the model is easily trainable utilising only five convolutional layers; in contrast, other state-of-the-art models use much deeper architectures [5–7]. Furthermore, the wide-kernel model can be scaled towards dimensionality changes of the data i.e., different numbers of sensors, and due to its compactness it is applicable in real-world settings (see [1] for the full model architecture).

Read the rest of the extended abstract in PDF.

Hybrid Approaches to Optimization and Machine Learning Methods (PDF)

Beatriz Flamia Azevedo, Ana Maria A. C. Rocha and Ana I. Pereira

Abstract:

This paper conducts a comprehensive literature review concerning hybrid techniques that combine optimization and machine learning approaches for clustering and classification problems. The aim is to identify the potential benefits of integrating these methods to address challenges in both fields. The paper outlines optimization and machine learning methods and provides a quantitative overview of publications since 1970. Additionally, it offers a detailed review of recent advancements in the last three years. The study includes a SWOT analysis of the top ten most cited algorithms from the collected database, examining their strengths and weaknesses as well as uncovering opportunities and threats explored through hybrid approaches. Through this research, the study highlights significant findings in the realm of hybrid methods for clustering and classification, showcasing how such integrations can enhance the shortcomings of individual techniques.

Sparse Self-Attention Guided Generative Adversarial Networks for Time-Series Generation (PDF)

Nourhan Ahmed and Lars Schmidt-Thieme

Abstract:

Remarkable progress has been achieved in generative modeling for time-series data with the introduction of Generative Adversarial Networks (GANs) [1]. GANs are neural networks that are meant to generate synthetic instances of data utilizing two neural networks, a generator and a discriminator, that operate against each other at the same time [1]. The generator learns to generate fake data to get the discriminator to classify its generated samples as authentic. The discriminator, on the other hand, attempts to distinguish between authentic and produced data. Finally, the generator could generate realistic data. GANs have demonstrated their ability to generate realistic data and have made remarkable progress in various tasks, such as the generation of time-series [4], images [5], and videos [3]. Particularly, a significant amount of work has utilized GANs based on Recurrent Neural Networks (RNNs) for time-series generation [4]. However, by carefully examining the generated samples from these models, we can observe that RNN-based GANs, such as LSTM GANs and gated recurrent GANs, cannot handle long sequences. Although RNN-based GANs can generate many realistic samples, there is still a difficulty in training due to exploding vanishing gradients and mode collapse

that limits their generation capability. In addition, these RNN-based GANs are typically designed for regular time-series data, and thus cannot maintain informative varying intervals properly, which is a major concern for generating time-series data.

Read the rest of the extended abstract in PDF.

Session: Optimization (Research VII)

AdaSub: Stochastic Optimization Using Second-Order Information in Low-Dimensional Subspaces (PDF)

João Victor Galvão da Mata and Martin Skovgaard Andersen

Abstract:

We introduce AdaSub, a stochastic optimization algorithm that computes a search direction based on second-order information in a low-dimensional subspace that is defined adaptively based on available current and past information. Compared to first-order methods, second-order methods exhibit better convergence characteristics, but the need to compute the Hessian matrix at each iteration results in excessive computational expenses, making them impractical. To address this issue, our approach enables the management of computational expenses and algorithm efficiency by enabling the selection of the subspace dimension for the search. Our code is freely available on GitHub, and our preliminary numerical results demonstrate that AdaSub surpasses popular stochastic optimizers in terms of time and number of iterations required to reach a given accuracy.

ISGP: Influence Maximization on Dynamic Social Networks Using Influence SubGraph Propagation (PDF)

Wan-Jhen Wu, Shiou-Chi Li and Jen-Wei Huang

Abstract:

Most previous research on influence maximization has focused on static social networks, despite the dynamic nature of networks in the real world. The computational cost imposed by recalculating results in response to dynamic changes precludes the use of conventional updating algorithms when dealing with large-scale networks. In this study, we developed a novel approach to estimating the influence of nodes through the creation of Influence SubGraphs. We also developed methods by which to update Influence SubGraphs to overcome the problem of influence maximization in dynamic social networks. Experiment results demonstrated the efficacy of the proposed scheme, in achieving influence propagation performance comparable to that of state-of-the-art methods with far lower memory requirements and far shorter execution times.

HyperTab: Hypernetwork Approach for Deep Learning on Small Tabular Datasets (PDF)

Witold Wydmański, Oleksii Bulenok and Marek Śmieja

Abstract:

Deep learning has achieved impressive performance in many domains, such as computer vision and natural language processing, but its advantage over classical shallow methods on tabular datasets remains questionable. It is especially challenging to surpass the performance of tree-like ensembles, such as XGBoost or Random Forests, on small-sized datasets (less than 1k samples). To tackle this challenge, we introduce HyperTab, a hypernetwork-based approach to solving small sample problems on tabular datasets. By combining the advantages of Random Forests and neural networks, HyperTab generates an ensemble of neural networks, where each target model is specialized to process a specific lower-dimensional view of the data. Since each view plays the role of data augmentation, we virtually increase the number of training samples while keeping the number of trainable parameters unchanged, which prevents model overfitting. We evaluated HyperTab on more than 40 tabular datasets of a varying number of samples and domains of origin and compared its performance with shallow and deep learning models representing the current state-of-the-art. We show that HyperTab consistently outranks other methods on small data (with statistically significant differences) and scores comparable to them on larger datasets.

Session: Journal III

DynamiSE: Dynamic Signed Network Embedding for Link Prediction (PDF)

Haiting Sun, Peng Tian, Yun Xiong, Yao Zhang, Yali Xiang, Xing Jia and Haofen Wang

Abstract:

In real-world scenarios, dynamic signed networks are ubiquitous where edges have positive and negative sign semantics and evolve over time. Encoding the dynamics and sign semantics of the network simultaneously is challenging. Moreover, over-smoothing is inevitably introduced by the learning of network dynamics. Targeting this gap, we propose Dynamic Signed Network Embedding (DynamiSE), which effectively integrates the balance theory and ordinary differential equation (ODE) into node representation learning to construct a deeper dynamic signed graph neural network and capture the complex sign semantics formed by the two types of edges.

PAF-Tracker: A Novel Pre-Frame Auxiliary and Fusion Visual Tracker (PDF)

Wei Liang, Derui Ding and Hui Yu

Abstract:

Relying on a large amount of data, recent object trackers achieve superior performance. However Siamese-like trackers expose considerable shortcomings in the case of brief occlusion. To address these shortages, the paper proposes a novel pre-frame auxiliary and fusion tracking framework. Within this framework, a retained variable is first introduced to avoid some additional twin branches while retaining the previously obtained deep features of the search frames. Based on such a variable, a pre-frame auxiliary module is constructed to establish the relationship between encoding features and the retained pre-frame information and a decoding fusion module is designed to fuse the generated similarity relationship. Moreover, the Efficient IoU (EIoU) loss is employed to increase the precision of predicted bounding boxes by adding three penalty terms for the differences in the center point, length, and width of the two bounding boxes. Finally, the superiority over state-of-the-art methods is verified by numerous tests on visual tracking benchmarks.

Entity Recognition Based on Heterogeneous Graph Reasoning of Visual Region and Text Candidate (PDF)

Xinzhi Wang, Nengjun Zhu, Jiahao Li, Yudong Chang and Zhennan Li

Abstract:

While significant progress has been made in recognizing entities from plain text, the exploration of entity recognition from multimodal data remains limited due to disparities in semantic representation. In light of this challenge, given the supportive nature of visual and text data, we propose a novel entity recognition model called Heterogeneous Graph Reasoning (HGR), leveraging the synergistic nature of visual and textual data. This is achieved through the utilization of the Vision Refine and Graph Cross Inference modules. In the Vision Refine module, semantically relevant objects hidden in the image are selected to aid in the text entity extraction. In the Graph Cross Inference module, cross-association inference between visual regions and textual entities is constructed through graph construction, heterogeneous graph fusion, visual region refinement and cross inference. Extensive experiments on four multimodal datasets are demonstrate the superiority of our model, when compared to the second-best state-of-the-art model.

Session: Emerging Problems in Disinformation (DISA)

Machine Learning-Based Android Malware Detection (PDF)

Carson Leung

Abstract:

The use of mobile phones, particularly smartphones, has been growing exponentially in recent times. From 2016 to 2021, smartphone users increased by more than 70%. With the increase in the popularity of smartphones, smartphones have become the prime target for criminal hackers. As a result, Android malware samples are coming to the market at an alarming rate. A study shows that there are more than 4 million malicious Android apps in the market, and each day around 11,000 new malwares add to this number. To combat this mass number of malware, we need a malware detection system that is efficient in detecting malicious Android apps. There are numerous existing malware detection systems, but most of them require countless features from both dynamic and static analysis. Thus, they are not scalable, lightweight, and efficient in detecting malware. Additionally, most studies that used limited features like only permission data, had done their research on much older dataset. Hence, there is a need for new research on this topic. In this paper, we build a permission-based malware detector for Android application with a new dataset and significantly less permissions. Initially, we used support vector machine (SVM) and all the extracted permission data as features to build our classification model. The model accuracy, precision, recall and F1 score were 97.41 percent, which is higher than the other state of the art similar approaches done on an older dataset. Next, we replicated this similar study with a few different machine learning algorithms: random forest, decision tree and logistic regression, and observed they all give similar results. However, tree-based algorithm performs a little better than the other algorithms. Finally, to achieve a lightweight malware detection system, we reduced the number of permissions or the features on a two-step process, and found only a slight difference in results. In the first step, even after reducing the number of permissions by about 94%, the accuracy dropped by only 2.7%. In the second step, we further reduced the number of features or permissions and observed the difference in results. We managed to prune to 9 permissions while maintaining accuracy of 93%, which is lower than technique mentioned in other literature to reduce features.

Model Stitching Algorithm for Fake News Detection Problem (PDF)

Rafał Kozik, Aleksandra Pawlicka, Marek Pawlicki and Michal Choras

Abstract:

Nowadays, we can see how social media networks are developing. We must accept the fact that the opinion of an expert is frequently just as valuable and crucial as that of a non-expert. It is feasible to see how traditional media is undergoing changes and processes that diminish the importance of the traditional "editing office" and place a growing focus on journalists' remote labour. As a result, social media has evolved into a component of national security since fake news and disinformation spread by nefarious individuals can influence readers and spark pointless debates on social issues that are inherently unimportant. This has a domino effect, instils dread in the populace, and eventually puts the security of the state in jeopardy. Recently, deep machine learning techniques have proven to be one of the technologies thought to be an effective way to combat the false news problem. However, due to shortages of labelled data, these methods often have poor model generalization capabilities when applied in real-world cases. In this paper, we address this problem by utilizing lightweight model stitching, which serves as a foundation for a hybrid method for fake news detection. Six distinct benchmark datasets have been used in our varied experiments. The outcomes are promising and pave the way for additional studies.

Towards Handling Bias in Intelligence Analysis with Twitter (PDF)

Alexandros Karakikes, Panagiotis Alexiadis, Theocharis Theocharopoulos, Nikolaos Skoulidas, Dimitris Spiliotopoulos and Konstantinos Kotis

Abstract:

Bias identification and mitigation in the Twitter ecosystem has been lately researched towards achieving a more efficient utilization of the application by different stakeholders and for a wide area of purposes. Among these stakeholders, intelligence services worldwide, collectively called the Intelligence Community (IC), tend to use Twitter, supplementary to their pre-existent disciplines, for monitoring areas of interest and identifying emerging social, political and security trends/threats. Over time, the IC has identified bias as the major obstacle in information analysis, thus it has developed scientific and empirical methods for bias mitigation, in parallel to those developed by the information and communication technology (ICT) and artificial intelligence (AI) community. As it becomes apparent, it is to both communities' interest to accurately trace bias and ideally eradicate or moderate its effects. In this paper we draw systemic parallels between Intelligence Analysis (IA) and Twitter Analytics (TA), comparatively examine existing bias mitigating methodologies to pinpoint similarities/dissimilarities, and utterly investigate the feasibility of adapting and adjusting methodologies from the first field to the latter. Furthermore, we propose a novel framework for AI-augmented bias mitigation in the IC. Finally, we propose methods and tools, already adapted by the ICT community, for efficiently supporting bias mitigation methodologies adapted by the IC.

A Continual Learning System with Self Domain Shift Adaptation for Fake News Detection (PDF)

Sebastián Basterrech, Andrzej Kasprzak, Jan Platos and Michal Wozniak

Abstract:

Detecting fake news is currently one of the critical challenges facing modern societies. The problem is particularly relevant, as disinformation is readily used for political warfare but can also cause significant harm to the health of citizens, such as by promoting false data on the harmfulness of selected therapies. One way to combat disinformation is to treat fake news detection as a machine learning task. This paper presents such an approach, which additionally addresses an important problem related to the non-stationarity characteristics of the fake news. We elaborated a stream data with the simulation of domain shift based on two popular benchmark datasets dedicated to the fake news classification problem (Kaggle Fake News and Constraint@AAAI2021–COVID19 Fake News Detection). The proposed learning system works in a Continual Learning (CL) framework and integrates a self-domain shift adaptation in a machine learning scheme. The method was built following state-of-the-art techniques, that includes Word2Vec as a feature extractor and the LSTM model as a classifier. The performance of the approach has been evaluated over the generated data stream. The convenience of our approach is showed in the results, where the accuracy gain with respect to a CL approach without domain adaptation is observed to be significant.

Combating Disinformation with Holistic Architecture, Neuro-symbolic AI and NLU Models (PDF)

Rafał Kozik, Wojciech Mazurczyk, Krzysztof Cabaj, Aleksandra Pawlicka, Marek Pawlicki and Michal Choras

Abstract:

It is important to realize that false news is more than just a deception. Sadly, it is impossible to confirm every bit of information we come across. A normal human impulse is to accept any information that looks sufficiently convincing, relevant, or exciting. In doing so, we often do not realize that we have just contributed to the misinformation of the community to which we belong. As a result, fake news happens to be our collective error. In this paper, we propose an architecture for combating the disinformation problem using a hybrid-based approach. We demonstrate our preliminary results on the health-related fake news dataset.

Session: Algorithms for Learning and Testing (Research VIII)

Tackling Model Mismatch with Mixup Regulated Test-Time Training (PDF)

Bochao Zhang, Rui Shao, Jingda Du, Pc Yuen and Wei Luo

Abstract:

Test-time training (TTT) is an emerging approach for addressing the problem of domain shift. In its framework, a test-time training phase is inserted between the training phase and the test phase. During the test-time training phase, the representation layers are adapted using an auxiliary task. Then the updated model will be used in the test phase. Although the idea is very intuitive, TTT does not demonstrate competitive performance compared with some other domain adaption methods. In this paper, we present both theoretical and empirical analyses to explain the subpar performance of TTT. In particular, we point out that TTT causes a new kind of problem, which we term as Model Mismatch. To address this problem of Model Mismatch, we analyse a simple yet effective method inspired by the idea of mixup in robust training. Such effectiveness is shown in the experimental results.

Natural Language Inference by Integrating Deep and Shallow Representations with Knowledge Distillation (PDF)

Pei-Chang Chen, Hao-Shang Ma and Jen-Wei Huang

Abstract:

Natural language understanding models often make use of surface patterns or idiosyncratic biases in a given dataset to make predictions pertaining to natural language inference (NLI) tasks. Unfortunately, this renders the resulting model vulnerable to out-of-distribution datasets to which the identified features are inapplicable, thereby leading to erroneous results. Many of the methods developed for out-of-distribution datasets have proven effective; however, they also tend to impose a trade-off in performance when applied to in-distribution datasets. In this paper, we use a teacher model providing knowledge for the student ensemble model as basic information for training. The student ensemble model then integrates information of deep and shallow representations to extend learning performance to a wide range of examples. The evaluation demonstrates that the proposed model outperformed state-of-the-art models when applied to in-distribution as well as out-of-distribution datasets.

Rapid and Scalable Bayesian AB Testing (PDF)

Srivas Chennu, Andrew Maher, Christian Pangerl, Subash Prabanantham, Jae Hyeon Bae, Jamie Martin and Bud Goswami

Abstract:

AB testing aids business operators with their decision making, and is considered the gold standard method for learning from data to improve digital user experiences. However, there is usually a gap between the requirements of practitioners, and the constraints imposed by the statistical hypothesis testing methodologies commonly used for analysis of AB tests. These include the lack of statistical power in multivariate designs with many factors, correlations between these factors, the need of sequential testing for early stopping, and the inability to pool knowledge from past tests. Here, we propose a solution that applies hierarchical Bayesian estimation to address the above limitations. In comparison to current sequential AB testing methodology, we increase statistical power by exploiting correlations between factors, enabling sequential testing and progressive early stopping, without incurring excessive false positive risk. We also demonstrate how this methodology can be extended to enable the extraction of composite global learnings from past AB tests, to accelerate future tests. We underpin our work with a solid theoretical framework that articulates the value of hierarchical estimation. We demonstrate its utility using both numerical simulations and a large set of real-world AB tests. Together, these results highlight the practical value of our approach for statistical inference in the technology industry.

Finite-Sample Bounds for Two-Distribution Hypothesis Tests (PDF)

Cynthia Hom, William Yik and George Montanez

Abstract:

With the rapid growth of large language models, big data, and malicious online attacks, it has become increasingly important to have tools for anomaly detection that can distinguish machine from human, fair from unfair, and dangerous from safe. Prior work has shown that two-distribution (specified complexity) hypothesis tests are useful tools for such tasks, aiding in detecting bias in datasets and providing artificial agents with the ability to recognize artifacts that are likely to have been designed by humans and pose a threat. However, existing work on two-distribution hypothesis tests requires exact values for the specification function, which can often be costly or impossible to compute. In this work, we prove novel finite-sample bounds that allow for two-distribution hypothesis tests with only estimates of required quantities, such as specification function values. Significantly, the resulting bounds do not require knowledge of the true distribution, distinguishing them from traditional p-values. We apply our bounds to detect student cheating on multiple-choice tests, as an example where the exact specification function is unknown. We additionally apply our results to detect representational bias in machine-learning datasets and provide artificial agents with intention perception, showing that our results are consistent with prior work despite only requiring a finite sample of the space. Finally, we discuss additional applications and provide guidance for those applying these bounds to their own work.

Session: Computational Imaging, Vision, Linguistics and Language (CIVIL I)

Adaptive Compressed Sensing for Real-Time Video Compression, Transmission, and Reconstruction (PDF)

Yaping Zhao, Qunsong Zeng and Edmund Lam

Abstract:

The real-time transmission of videos with both high resolution and high frame rate is challenging, due to the limited storage space and significant communication overhead. To meet the real-time requirement, these issues are usually tackled by video quality reduction, which compromises the user experience. While previous methods, such as video compressed sensing, have attempted to address these issues, they often employ a fixed compression rate without considering the varying channel gain and do not adequately address the real-time transmission requirements. To mitigate these shortcomings, we propose an adaptive compressed sensing framework that optimizes the compression rate based on the channel state. This approach equivalently optimizes the video quality while ensuring real-time transmission by reducing communication overhead and thus latency. The feasibility and performance of our method are validated and discussed through extensive experiments on both classic and custom datasets.

A CNN-Transformer Hybrid Network for Multi-scale object detection (PDF)

Jianhong Wu and Yingdong Ma

Abstract:

Recently, Transformer-based methods have been the main framework in various computer vision tasks. Vision transformers achieve object detection based on sequence of visual tokens, lacking the ability of extracting local context and dealing with scale variance. To tackle this problem, we propose a hybrid CNN-transformer model with multiple dual-branch transformer blocks in which transformer branch captures global dependences and the CNN branch enhances local context. As convolution branch and transformer branch pay attention to different-level information, we combine the output features of the CNN branch and transformer branch with adaptive weights calculated from the visual content. Moreover, instead of detecting objects from transformer outputs directly, we introduce a feature aggregation module to fuse different levels features and construct feature pyramid based upon these multi-level features. The proposed feature aggregation module alleviates semantic gap between high-level and low-level features. Experimental results on the MS COCO dataset show that our method significantly improves the performance of multi-scale object detection.

Searching Images in a Web Archive (PDF)

André Mourão and Daniel Gomes

Abstract:

This article presents the research and development of a large-scale image search system applied to launch a word-wide innovative service that enables searching billions of historical images archived from the web since the 1990s. Contributions of this work were applied to enhance the Arquivo.pt web archive with an image-search service where users submit text queries, through a web user interface or an API, and immediately receive a list of historical web-archived images. However, supporting image search over web archives raised new challenges. The volume of data to be processed was big and heterogeneous, summing over 530TB of historical web data published since the early days of the web. The main contributions of this work are a toolkit of algorithms that extracts textual metadata to describe web-archived images, a system architecture and workflow to index large amounts of web-archived images considering their specific temporal features and a ranking algorithm to order image-search results by relevance. This research was applied to launch an enhanced image-search service that is publicly available since March 2021. All the developed software is fully available as free open-source software.

ScaleFace: Uncertainty-aware Deep Metric Learning (PDF)

Roman Kail, Kirill Fedyanin, Nikita Muravev, Alexey Zaytsev and Maxim Panov

Abstract:

The performance of modern deep learning-based systems dramatically depends on the quality of input objects. For example, face recognition quality is lower for blurry or corrupted inputs. Moreover, it is difficult to predict the influence of input quality on the resulting accuracy in more complex scenarios. We propose a deep metric learning framework that allows for direct estimation of the uncertainty with almost no additional computational cost. The developed *ScaleFace* algorithm uses trainable scale values that modify similarities in the space of embeddings. These input-dependent scale values represent a measure of confidence in the recognition result, thereby providing provably reasonable uncertainty estimation. We present results from comprehensive experiments on open-set classification tasks, including face recognition, which demonstrate the superior performance of ScaleFace compared to other uncertainty-aware face recognition approaches. We also extend our study to the task of text-to-image retrieval, showing that the proposed approach outperforms competitors by significant margins.

Session: Smart City Data Analytics (SmartCities I)

Empowering Urban Connectivity in Smart Cities using Federated Intrusion Detection (PDF)

Youcef Djenouri and Ahmed Nabil Belbachir

Abstract:

The advent of transformative technologies such as the Internet of Things (IoT) has brought forth significant advancements in various sectors like smart cities, fintech, learning, and healthcare, as well as revolutionized online activities. The IoT has facilitated widespread connectivity by interconnecting numerous objects and services, but it has also made IoT and cloud infrastructures susceptible to cyberattacks, making cybersecurity a paramount concern, particularly for the development of reliable IoT systems, especially those powering smart city networks. In this research endeavor, we embark on exploring a cutting-edge pipeline that amalgamates federated deep learning with a trusted authority approach to tackle the intricate challenges associated with intrusion detection in smart city networks. To identify anomalies and intrusions effectively within the network, we devise an improved LSTM (Long Short-Term Memory) model. Additionally, we propose an intelligent swarm optimization solution to address dimensionality reduction concerns. Thorough evaluations of our federated learning-based approach are conducted, and these are juxtaposed with several basic approaches, utilizing the renowned NSL-KDD dataset. Encouragingly, our findings reveal that the proposed framework remarkably outperforms the baseline solutions, particularly when dealing with datasets containing a substantial volume of transactions. Furthermore, our method ensures robust data security for the model, as it becomes the pioneering endeavor to incorporate the principle of trusted authority into the realm of federated learning for the management of smart city networks.

Recycling of Generic ImageNet-trained Models for Smart-city Applications (PDF)

Katarzyna Filus and Joanna Domanska

Abstract:

Convolutional Neural Networks (CNNs) enabled breakthroughs in computer vision. They are also used in different domains of smart cities to process and analyze large amounts of image data, which is crucial for intelligent decision-making. CNNs trained on large-scale datasets such as ImageNet, possess remarkable image classification abilities. As it takes a lot of time and computational resources to train models on such datasets, we propose a solution to reuse these models in the area of smart-cities. We call our approach the model recycling, as it uses the existing versatile knowledge of ImageNet-trained networks for resource conservation and faster deployment of applications in the smart-city domains. For that purpose, we utilize the semantic connections between ImageNet categories to determine the sets of classes that can be used to create specialized classification models for smart transportation, shopping and education with no additional training. We present a methodology to extract such specialized models from generic CNNs trained on ImageNet. Such models can be used at low budget to create solutions for automatic data annotation and in interactive applications. They can also be used as a starting point for fine-tuning. We also present 2 strategies of creating ensembles with these specialized models for better overall and per-class accuracy. Our general methodology can be also used in other domains outside the smart cities.

Incremental Targeted Mining in Sequences (PDF)

Kaixia Hu, Wensheng Gan, Gengsen Huang, Guoting Chen and Jerry Chun-Wei Lin

Abstract:

High utility sequential pattern mining (HUSPM) is a critical research topic in data analytics (e.g., smart-city technologies), which takes into consideration three pivotal factors of data: timestamp, internal quantization, and external utility. Recently, a query-enabled HUSPM approach has been proposed, which aims to discover patterns based on a query sequence. However, this approach only works on static data and does not solve the tasks well under dynamic data. When the data is updated, it needs to restart the mining process, which leads to a lot of duplicate calculations and resource consumption. In the paper, to address the mining task of increasing sequence data over time, we develop an Incremental Targeted HUSPM algorithm called ITUS. A tighter upper bound called tight extension sequence utility (TESU) is proposed to determine key candidates, which can avoid the generation of unpromising patterns. By using TESU, a target candidate pattern tree (TCP-tree) is utilized to record the sequence information, and several efficient strategies are implemented to incrementally update the tree. Finally, we extensively evaluate our proposed algorithm on both real-world and synthetic datasets. The experimental results clearly demonstrate that not only does the novel algorithm guarantee the accuracy of the results after multiple database updates, but it also achieves higher efficiency than the baseline approach.

Price Prediction of Digital Currencies Using Machine Learning (PDF)

Ashutosh Dhar Dwivedi, Subhrangshu Adhikary, Subhayu Dutta and Jens Myrup Pedersen

Abstract:

Cryptocurrencies have gained immense significance and popularity in recent times. With thousands of digital currencies available, selecting the right one can be challenging for users. In the financial sector, accurately predicting future prices is crucial for profitable investments in digital currencies. However, price prediction in this realm poses unique challenges, as it lacks physical goods or services as the basis, unlike stock prices. Machine learning emerges as a pivotal tool for addressing this challenge and plays a vital role in price prediction. This research analyzes five prominent currencies – Monero, Bitcoin, Ethereum, IOTA, and Zcash –

employing five models: SVR, LRG, Huber, RANSAC, MLP, and AdaBoost. The experimental results demonstrate promising outcomes, showcasing the ability to predict digital currency prices with an impressive R2 score of 1.0 for specific machine learning algorithms. This advancement opens new avenues for informed decision-making and profitable ventures in the dynamic world of digital currencies.

ZigBee Network for AGV Communication in Industrial Environments (PDF)

Jarosław Flak, Tomasz Skowron, Rafał Cupek, Marcin Fojcik, Dariusz Caban and Adam Domański

Abstract:

Automated Guided Vehicles (AGVs) are a key component of many modern industrial systems. AGVs are supposed to communicate with each other in real time using wireless networks. In this article, the advantages and disadvantages of the ZigBee wireless network related to the control of AGVs are considered. We analyze the performance of the ZigBee network programmed with both C# and Python libraries to control ZigBee devices. The throughput and signal strength are presented and discussed depending on the transmission speed of the serial port, the payload size, and the presence and distance from the obstacles. The results of the experiments show the effective values of these parameters, the methods of using C# and Python, and the reliability of the throughput up to a certain point in network devices.

Session: Computational Imaging, Vision, Linguistics and Language (CIVIL II)

YOLO-based Object Detection in Panoramic Images of Smart Buildings (PDF)

Sebastian Pokuciński and Dariusz Mrozek

Abstract:

Collecting and analyzing spherical images is one of the crucial fields supporting the digitization of indoor environments, like houses, apartments, offices, or factories, and creating virtual tours of smart buildings. Such images also constitute an important feed for smart indoor devices, like autonomous vacuum cleaners, intelligent production lines, or assistive droids. However, smart devices must correctly identify internal objects on high-resolution spherical images, usually heavily distorted and made with variable lighting conditions. Moreover, the recognition task must be relatively accurate and take place onboard the device, as data transmission to larger analysis centers or workstations does not allow for real-time operation. In this paper, we compare two object detectors from the YOLO family (YOLOv5 and YOLOv8) on the publicly available dataset adjusted to a newer annotation representation that allows for adapting YOLO detectors to spherical images. We verify the effectiveness of these two families of detectors for varying sizes of object detection models, image sizes, and training batch sizes. Our research proves that YOLO models can successfully detect most indoor objects, and their effectiveness is comparable to dedicated detectors having much higher computational complexity.

Solving Inverse Problems in Compressive Imaging with Score-Based Generative Models (PDF)

Zhen Yuen Chong, Yaping Zhao, Zhongrui Wang and Edmund Lam

Abstract:

Snapshot Compressive Imaging (SCI) is a technique for capturing high-dimensional data through snapshot measurements using a two-dimensional (2D) detector. This approach is accomplished via coded aperture compressive temporal imaging (CACTI), which involves applying a temporally variant mask to spatially encode each sequential signal before aggregating the encoded information into a single compressed measurement. The objective of our work is to develop algorithms capable of reconstructing each video frame as a 3D data cube from its 2D measurement. To achieve this goal, we introduce multiple approaches that utilize unconditional and pre-trained 2D score models for video frame reconstruction. Our method involves modeling both the forward perturbation process of the data distribution and its reverse process as stochastic differential equations (SDEs). We also employ score-based deep learning models to estimate the scores of the data distribution across different time steps. Differing from many applications, our sampling process relies on the observed measurement, which directly corresponds to pixel values, rather than class labels. We demonstrate that employing traditional score-based generative methods with 2D score models in SCI, or integrating them into the plug-and-play framework as a deep generative prior, presents challenges. Furthermore, we propose ideas to address these limitations for future research.

All Translation Tools Are Not Equal: Investigating the Quality of Language Translation for Forced Migration (PDF)

Ameeta Agrawal, Lisa Singh, Elizabeth Jacobs, Yaguang Liu, Gwyneth Dunlevy, Rhitabrat Pokharel and Varun Uppala

Abstract:
As the volume and complexity of forced movement continues to grow, there is an urgent need to use new data sources to better understand emerging crises. Organic sources, like social media and newspapers, can offer insights in near real time when administrative data are unavailable for timely and detailed analysis. However, to flexibly switch to different contexts, we need the ability to contextualize the drivers of movement for different locations and languages. Recent advances in natural language processing and specifically, neural machine translation, have shown impressive results on standard benchmark datasets for well-studied language pairs. However, the effectiveness of these models in a real-world scenario remains less known. To advance our understanding of real-world, contextual translation, we systematically study the performance of multiple widely used off-the-shelf machine translation tools using words associated with drivers of forced movement in both high- and low-resource languages. Our empirical results suggest significant variation between the performance of these machine translation tools in terms of accuracy and efficiency, highlighting a problem that must be faced by those conducting migration research using multilingual contexts. We conclude by suggesting strategies for obtaining reasonable translations from off-the-shelf language tools.

Session: Graph Data Science and Applications (GraDSI I)

JAMES: Normalizing Job Titles with Multi-Aspect Graph Embeddings and Reasoning (PDF)

Michiharu Yamashita, Jia Tracy Shen, Hamoon Ekhtiari, Thanh Tran and Dongwon Lee

Abstract:
In online job marketplaces, it is important to establish a well-defined job title taxonomy for various downstream tasks (e.g., job recommendation, users' career analysis, and turnover prediction). Job Title Normalization (jtn) is such a cleaning step to classify user-created non-standard job titles into normalized ones. However, solving the jtn problem is non-trivial with challenges: (1) semantic similarity of different job titles, (2) non-normalized user-created job titles, and (3) large-scale and long-tailed job titles in real-world applications. To this end, we propose a novel solution, named JAMES, that constructs three unique embeddings (i.e., graph, contextual, and syntactic) of a target job title to effectively capture its various traits. We further propose a multi-aspect co-attention mechanism to attentively combine these embeddings, and employ neural logical reasoning representations to collaboratively estimate similarities between messy job titles and normalized job titles in a reasoning space. To evaluate JAMES, we conduct comprehensive experiments against ten competing models on a large-scale real-world dataset with over 350,000 job titles. Our experimental results show that JAMES significantly outperforms the best baseline by 10.06% in Precision@10 and by 17.52% in NDCG@10, respectively. To further facilitate the acquisition of normalized job titles for job-domain applications, we will release the codebase and the public API of JAMES at: https://tinyurl.com/james-job-title-mapping.

Unfolding Temporal Networks through Statistically Significant Graph Evolution Rules (PDF)

Alessia Galdeman, Tommaso Locatelli, Matteo Zignani and Sabrina Gaito

Abstract:
Understanding and extracting knowledge from temporal networks is crucial to understand their dynamic nature and gain insights into their evolutionary characteristics. Existing approaches to network growth often rely on single-parameterized mechanisms, neglecting the diverse and heterogeneous behaviors observed in contemporary techno-social networks. To overcome this limitation, methods based on graph evolution rules (GER) mining have proven promising. GERs capture interpretable patterns describing the transformation of a small subgraph into a new subgraph, providing valuable insights into evolutionary behaviors. However, current approaches primarily focus on estimating subgraph frequency, neglecting the evaluation of rule significance. To address this gap, we propose a tailored null model integrated into the GERM algorithm, the first and most stable graph evolution rule mining method. Our null model preserves the graph's static structure while shuffling timestamps, maintaining temporal distribution, and introducing randomness to event sequences. By employing a z-score test, we identify statistically significant rules deviating from the null model. We evaluate our methodology on three temporal networks representing co-authorship and mutual online message exchanges. Our results demonstrate that the introduction of the null model affects the evaluation and interpretation of

identified rules, revealing the prevalence of under-represented rules and suggesting that temporal factors and other mechanisms may impede or facilitate evolutionary paths. These findings provide deeper insights into the dynamics and mechanisms driving temporal networks, highlighting the importance of assessing the significance of the evolution patterns in understanding network evolution.

Prediction of Future Nation-initiated Cyber Attacks from News-based Political Event Graph (PDF)

Bishal Lakha, Jason Duran, Edoardo Serra and Francesca Spezzano

Abstract:

In the world of cyber defense, anticipating potential attacks or any increase in risk of attacks is one of the most advantageous pieces of knowledge one can have. However, little research has been done in examining the larger geopolitical environment and using data sources available at the geopolitical level to predict cyberattacks in advance. To this end, we combine the use of a geopolitical conflict dataset, ICEWS, in combination with a cyberattack dataset from the Council on Foreign Relations to determine if we can predict cyberattacks targeting a given nation. We present a novel approach to identify periods of increased likelihood of cyberattacks at the country, regional, and global levels. The approach involves creating a news-based political event graph, generating vectoral representations of the graph using the SIR-GN structural iterative representation learning approach, and applying novelty detection models to predict future nation-initiated attacks. The proposed approach outperforms existing baselines for majority of cases in terms of F1-score, demonstrating its effectiveness in predicting cyberattacks.

Session: Smart City Data Analytics (SmartCities II)

Automated Detection of Trajectory Groups Based on SNN-Clustering and Relevant Frequent Itemsets (PDF)

Friedemann Schwenkreis

Abstract:

Classification has been proposed for the automated detection of similarity groups in spatio-temporal data. However, recent approaches have introduced clustering based solutions to avoid the huge overhead for the manual classification of training and test data. This paper presents a combination of shared nearest-neighbor clustering and an adapted search for frequent itemsets to not only find similarity groups in sets of trajectories called team moves but also clusters of similar individual trajectories. Dynamic Time Warping is introduced as the underlying notion of trajectory distance on which the notion of trajectory similarity will be defined. Since the search for frequent itemsets is used to find similarity groups of team moves, an explicit distance criterion for team moves can be avoided. However, a notion of relevance is introduced that allows to distinguish trajectories with an impact on team moves from others. In addition, the paper will introduce enhanced quality indexes for shared nearest neighbor based trajectory clustering that allow to compare parameter settings to find the optimal clustering solution for a given problem.

Object-aware Multi-criteria Decision-Making Approach Using Heuristic data-driven Theory for Intelligent Transportation Systems (PDF)

M S Mekala, Elyad Eyad and Gm Srivastava

Abstract:

Sharing up-to-date information about the surrounding measured by On-Board Units (OBUs) and Roadside Units (RSUs) is crucial in accomplishing traffic efficiency and pedestrians safety towards Intelligent Transportation Systems (ITS). Transferring measured data demands ≥ 10Gbit/s transfer rate and ≥ 1GHz bandwidth though the data is lost due to unusual data transfer size and impaired line of sight (LOS) propagation. Most existing models concentrated on resource optimization instead of measured data optimization. Subsequently, RSU-LiDARs have become increasingly popular in addressing object detection, mapping and resource optimization issues of Edge-based Software-Defined Vehicular Orchestration (ESDVO). In this regard, we design a two-step data-driven optimization approach called Object-aware Multi-criteria Decision-Making (OMDM) approach. First, the surroundings-measured data by RSUs and OBUs is processed by cropping object-enabled frames using YoLo and FRCNN at RSU. The cropped data likely share over the environment based on the RSU Computation-Communication method. Second, selecting the potential vehicle/device is treated as an NP-hard problem that shares information over the network for effective path trajectory and stores the cosine data at the fog server for end-user accessibility. In addition, we use a non-linear programming multi-tenancy heuristic method to improve resource utilization rates based on device preference predictions (Like detection accuracy and bounding box tracking) which elaborately concentrate in future work. The simulation results agree with the targeted effectiveness of our approach, i.e., mAP(≥ 71%) with processing delay (≤ 3.5 × 106 bits/slot), and transfer delay (≤ 38ms). Our simulation results indicate that our approach is highly effective.

Predicting Conflict Zones on Terrestrial Routes of Automated Guided Vehicles with Fuzzy Querying on Apache Kafka (PDF)

Bozena Malysiak-Mrozek, Stanisław Kozielski and Dariusz Mrozek

Abstract:

In today's world, smart factories are a coexisting element of smarticizing cities. Smart manufacturing of today relies on the automation of many component tasks of the production process. Automated guided vehicles (AGVs) that transport materials on the production lines are important elements of this automation. Appropriate management of a fleet of AGVs requires avoiding collisions. However, prediction and early detection of approaching collision points on the transportation routes not only prevent collisions but also enables adjusting the AGV operation and improving its flow. In this paper, we demonstrate the use of fuzzy sets and linguistic variables in collision prevention by processing AGV data streams with Apache Kafka. We extend the capabilities of Apache Kafka and ksqlDB towards fuzzy stream processing and use fuzzy KSQL queries to predict collisions. Our experiments prove that fuzzy querying against AGV data streams does not consume much time and computational resources, and we can successfully avoid collisions by predicting future positions of the AGV for various densities of data streams and widths of time windows.

Low-Cost Gunshot Detection System with Localization for Community Based Violence Interruption (PDF)

Isaac Manring, James Hill, Paul Brantingham, George Mohler, Thomas Williams and Bruce White

Abstract:

There is growing interest in U.S. cities to shift resources towards community-led solutions to crime and disorder. However, there is a simultaneous need to provide community organizations with access to real-time data to facilitate decision making, to which only the police normally have access. In this work we present a low-cost gunshot detection system with localization that has been developed for community-based violence interruption. The distributed real-time gunshot detection sensor network is linked to a mobile phone-based alert and tasking system for exclusive use by civilian gang interventionists. Here we present details on the system architecture and gunshot detection model, which consists of an Audio Spectrogram Transformer (AST) neural network. We then combine gradient maps of the input to the AST for time of arrival identification with a Bayesian maximum a posteriori estimation procedure to identify the location of gunshots. We conduct several experiments using simulated data, open data from the commercial ShotSpotter detection system in Pittsburgh, and data collected using our devices during live-fire experiments at the Indianapolis Metropolitan Police Department (IMPD) gun firing range. We then discuss potential applications of the system and directions for future research.

Session: Data Science for Social and Behavioral Analytics (DSSBA)

Enhanced Mining of High Utility Patterns from Streams of Dynamic Profit (PDF)

Carson Leung

Abstract:

Frequent pattern mining has been extended to the mining of other useful patterns. These include high-utility patterns. Many traditional high-utility mining algorithms focus on algorithmic efficiency when mining high-utility patterns from static databases. These algorithms rely on an assumption that the unit utility for a given item is a constant. However, as we are living in dynamic world where the unit utility (external unit profit) may change over time, such an assumption may not truly reflect reality in the real world. However, to the best of our knowledge, not a lot of works were done on mining dynamic profit from data streams yet. The emergence of big data has led to some performance challenges such that proper big data management techniques are needed for knowledge discovery from dynamic data streams. Traditional static data mining algorithms cannot directly apply to dynamic data. Furthermore, information in the data stream might not be uniformly distributed so it introduces extra challenges to process the data. Using big data stream processing platforms is necessary when mining real-world data stream. Leveraging the big data processing framework requires having scalable algorithms. In this paper, we present an enhanced high-utility data stream algorithm—called EHUI-Stream—to speed up the execution time and reduce memory usage. Utilizing our proposed algorithm, the data stream

mining performance is expected to be further enhanced against both real-world datasets and synthetic datasets. Evaluation results on real-life data demonstrate the effectiveness of our platform in scalable high-utility pattern mining for dynamic profit from data streams for social and behavioral analytics.

Towards Contiguous Sequences in Uncertain Data (PDF)

Zefeng Chen, Wensheng Gan, Gengsen Huang, Yanxin Zheng and Philip S. Yu

Abstract:

In data mining, high-utility sequential pattern mining (HUSPM) focuses more on the specific values of items than on their frequency, making it more practical in real-life scenarios. HUSPM with the contiguous constraint can be used to solve some applications requiring the sequence elements to occur consecutively. Due to device, environment, privacy issues, and other factors, the data is often not accurate, and traditional algorithms for mining high utility continuous sequence patterns (HUCSPs) do not perform well in handling uncertain data. To address this challenge, this paper presents a new algorithm named uncertain utility-driven contiguous pattern mining (UUCPM), which can discover HUCSPs efficiently and correctly. The algorithm is designed to obtain results from sequence data with uncertain probabilities set on the item level. Two tighter upper bounds on utility and corresponding pruning strategies are also proposed, which can effectively process and reduce the number of candidate patterns generated during pattern mining, thereby improving the performance of the mining process. Through extensive experiments, the proposed UUCPM algorithm has been verified for accuracy and performance, demonstrating its advanced properties.

Emotion-based Dynamic Difficulty Adjustment in Video Games (PDF)

Krzysztof Kutt, Łukasz Ściga and Grzegorz J. Nalepa

Abstract:

Current review papers in the area of Affective Computing and Affective Gaming point to a number of issues with using their methods in out-of-the-lab scenarios, making them virtually impossible to be deployed. On the contrary, we present a game that serves as a proof-of-concept designed to demonstrate that—being aware of all the limitations and addressing them accordingly—it is possible to create a product that works in-the-wild. A key contribution is the development of a dynamic game adaptation algorithm based on the real-time analysis of emotions from facial expressions. The obtained results are promising, indicating the success in delivering a good game experience.

Session: Graph Data Science and Applications (GraDSI II)

Leveraging patient similarities via graph neural networks to predict phenotypes from temporal data (PDF)

Dimitrios Proios, Anthony Yazdani, Alban Bornet, Julien Ehrsam, Islem Rekik and Douglas Teodoro

Abstract:

Several machine learning approaches have been proposed to automatically derive clinical phenotypes from patient data. Nevertheless, methods leveraging similarity-based patient networks remain underexplored for temporal data. In this work, we propose a graph neural network (GNN) model that learns patient representation using different network configurations and feature modes. To explore the sequential nature of time series, features were extracted using a recurrent neural network (RNN) and embedded using information from the network structure via the GNN. Our method improves upon statistical and RNN baselines, with performance boosts up to 1% and 22% accuracy in the inductive and transductive settings, respectively. We also show that network configurations significantly impact performance in the transductive learning setting. Thus, automated phenotyping models based on GNNs could be used to support phenotype-based clinical research and ultimately for personalized clinical decision support.

Enhancing Recommendation Systems with Hybrid Manifold Regularized Knowledge Graph (PDF)

Giang Ngo and Nhi Vo

Abstract:

Recent advances in graph neural networks have motivated the use of structured information in the form of a knowledge graph in recommendation systems. The most important challenges of real-world recommendation

problems are the sparsity issues in which labeled user-item interaction data are sparse compared to the size of the knowledge graph. In this work, we propose a graph-based semi-supervised learning method namely Hybrid Manifold Regularized Knowledge Graph (HMR-KG), to tackle this problem. HMR-KG is an attentive knowledge graph neural network with a lightweight version of manifold regularization to enforce smoothness on the mapping function. Translation-based pre-trained embeddings are also used as initialization for the attentive knowledge graph neural networks. The method is evaluated using three public datasets. Experimental results show that our method not only outperforms state-of-the-art baselines but also yields more consistent results in scenarios with sparsely labeled data. In addition, our lightweight implementation successfully approximates the manifold regularization loss with substantially smaller time complexity.

EvoAlign: A Continual Learning Framework for Aligning Evolving Networks (PDF)

Shruti Saxena and Joydeep Chandra

Abstract:

Network alignment, the task of identifying the same entities across multiple networks, has recently gained immense popularity in industry and academia. However, most existing alignment methods consider networks static, merely a snapshot of the time-evolving real networks where the nodes, links, and attributes are bound to change over time. The existing methods cannot capture these intrinsic dynamic network changes and cannot be applied directly to the evolving real-network scenario. Even extending these static methods to update their model parameters by retraining dynamically suffers from high computational complexity or simple sequential training suffers from compromised predictions due to distribution drifts in the networks. Moreover, dealing with a pair of evolving networks poses extra challenges of their domain differences and different evolution rates and patterns. Hence to overcome these challenges, we propose EvoAlign, an end-to-end information replay-based continual learning framework built upon Graph Neural Networks (GNNs) for the alignment of evolving networks. EvoAlign employs the concept of uncertainty in model predictions to identify and address two key aspects: detecting new patterns and preserving historical patterns. It also uses a novel shift-induced regularizer to handle distribution drift and domain differences in evolving networks. We empirically show that our method outperforms the existing state-of-the-art alignment methods on three real datasets.

Session: Learning from Temporal Data (LfTD)

1NN-DTW ARIMA LSTM: A New Ensemble for Forecasting Multi-domain/Multi-context Time Series (PDF)

Hadi Fanaee-T

Abstract:

The measurements of several sensors of multiple machines are collected in real-time at Alfa Laval’s IIoT platform. It is of great interest to make an accurate forecast for upcoming events, such as machine failure or sensor faults, which are critical enablers for predictive maintenance. However, some machines have little historical data, making forecasting challenging for them. Can we use data from other devices to make a forecast on a newly installed machine or improve the forecasting performance on different machines? This is the central question I try to answer in this research. However, another dimension of complexity makes the problem even more difficult. The devices have some differences. They can vary in version, application, and configuration. In addition, at each instant, the machine can have different modes and user-defined parameters that can be changed anytime by the operator or machine auto-control. All these factors make the forecasting problem extremely challenging, so it does not fit into a general computational framework. I propose a new multi-paradigm ensemble framework, aware of context and domain, that significantly improves the popular ensemble methods and non-ensemble state-of-the-art such as ARIMA and LSTM. The improvement mostly comes from adding a new forecasting paradigm based on multi-domain 1NN-DTW (one-nearest neighbor with dynamic time warping as a similarity measure) to the ensemble architecture, which creates a surprising harmony with ARIMA. 1NN-DTW is popular in time series classification, but its application is being investigated for the first time in a heterogeneous ensemble. The other novel component is a meta-selection mechanism for aggregating forecasts, which unexpectedly works better than the widely-used meta-learning approach.

Online Explainable Model Selection for Time Series Forecasting (PDF)

Amal Saadallah

Abstract:

Several machine learning models have been used to tackle time series forecasting. However, it is generally accepted that none of them is universally valid for every application and over time. Therefore, adequate and adaptive real-time model selection is often required to cope with the time-evolving nature of time series and the fact that models have specific Regions of Competence (RoCs) across the time series. In this paper, we perform an online single model selection for time series forecasting by using an adaptive clustering method to compute the RoCs of candidate models. This method can be extended to ensemble base models selection by combining clustering with a rank-based approach. In this framework, the appropriate model(s) is selected online, and the RoCs responsible for model selection update is done adaptively in an informed manner following concept drift detection in the RoCs' structure. Moreover, the computed RoCs can be used to provide suitable explanations for the reason for selecting certain model(s) in a certain time interval or instant. Since the RoCs are computed independently of the family of forecasting models in question, the explanations we provide are model-agnostic. An extensive empirical study on various real-world datasets shows that our method achieves excellent or on-par results compared to state-of-the-art approaches and various baselines.

LITE: Light Inception with boosTing tEchniques for Time Series Classification (PDF)

Ali Ismail-Fawaz, Maxime Devanne, Stefano Berretti, Jonathan Weber and Germain Forestier

Abstract:

Deep learning models have been shown to be a powerful solution for Time Series Classification (TSC). State-of-the-art architectures, while conducting promising results on the UCR archive, present a high number of trainable parameters. This can lead to long training with a high CO2, Power consumption and possible increase in the number of FLoat-point Operation Per Second (FLOPS). In this paper, we present a new architecture for TSC, the Light Inception with boosTing tEchnique (LITE) with only 2.34% of the state-of-the-art model InceptionTime's number of parameters, while preserving performance. This architecture, with only 9,814 trainable parameters due to the usage of DepthWise Separable Convolutions (DWSC), is boosted by three techniques: multiplexing, custom filters, and dilated convolution. The LITE architecture, trained on the UCR, is 2.78 times faster than InceptionTime and consumes 2.79 times less CO2 and Power.

DSAA’23 Conference Committees

General Chairs

Yannis Manolopoulos, Open University of Cyprus, Cyprus
Zhi-Hua Zhou, Nanjing University, China

PC Chairs, Research Track

Guoliang Li, Tsinghua University, China
Timos Sellis, Archimedes/Athena RC, Greece

PC Chairs, Applications Track

Minos Garofalakis, Technical University of Crete, Greece
Takashi Washio, Osaka University, Japan

PC Chairs, Industrial Track

Peter Triantafillou, Warwick University, UK
Athena Vakali, Aristotle University of Thessaloniki, Greece

PC Chairs, Journal Track

Bin Yang, Aalborg University, Denmark
Feida Zhu, Singapore Management University, Singapore

Award Chair

Torben Bach Pedersen, Aalborg University, Denmark

Special Sessions Chairs

Vagelis Papalexakis, University of California Riverside, USA
Grant Scott, University of Missouri, USA

Trends and Controversy Chair

Charalampos Tsourakakis, Boston University, USA

Tutorial Chair

Themis Palpanas, Université Paris Cité, France

Data Science Competition Chair

Apostolos Papadopoulos, Aristotle University of Thessaloniki, Greece

Local Arrangements Chair

Anastasios Gounaris, Aristotle University of Thessaloniki, Greece

Publicity/Social Media Chairs

David Anastasiu, Santa Clara University, USA
Qiang He, Swinburne University of Technology, Australia
Francesca Pratesi, CNR, Italy
Salma Sassi, University of Pau and the Adour Region, France

Publication Chair

Karam Bou-Chaaya, EXPLEO GROUP, France

Treasurer

Richard Chbeir, Université de Pau et des Pays de l’Adour, France

Webmaster

Elie Chicha, Université de Pau et des Pays de l’Adour, France