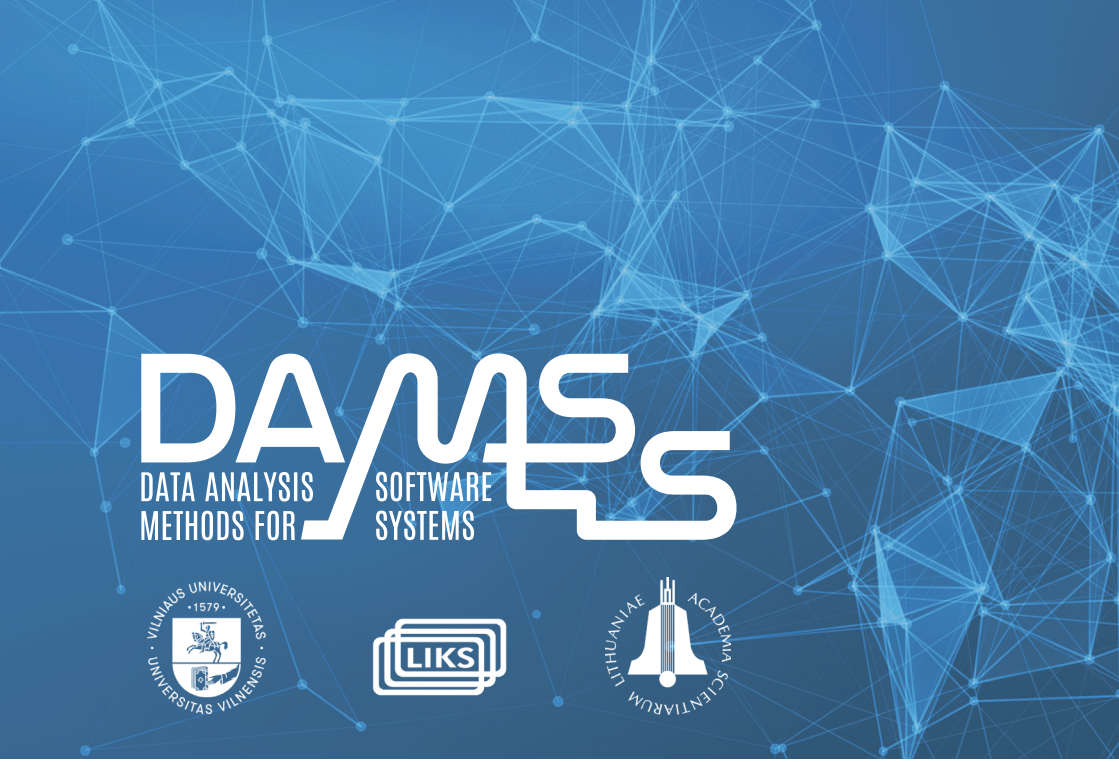

14th Conference on

DATA ANALYSIS METHODS for Software Systems

November 30 – December 2, 2023

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

LITHUANIAN COMPUTER SOCIETY
VILNIUS UNIVERSITY INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES
LITHUANIAN ACADEMY OF SCIENCES

14th Conference on

DATA ANALYSIS METHODS
for Software Systems

November 30 – December 2, 2023

Druskininkai, Lithuania, Hotel "Europa Royale"
https://www.mii.lt/DAMSS

VILNIUS UNIVERSITY PRESS
Vilnius, 2023

Co-Chairmen:

Prof. Gintautas Dzemyda (Vilnius University, Lithuanian Academy of Sciences)
Dr. Saulius Maskeliūnas (Lithuanian Computer Society)

Programme Committee:

Dr. Jolita Bernatavičienė (Lithuania)
Prof. Juris Borzovs (Latvia)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Janusz Kacprzyk (Poland)
Prof. Ignacy Kaliszewski (Poland)
Prof. Bożena Kostek (Poland)
Prof. Tomas Krilavičius (Lithuania)
Prof. Olga Kurasova (Lithuania)
Assoc. Prof. Tatiana Tchemisova (Portugal)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

Dr. Jolita Bernatavičienė
Prof. Olga Kurasova
Assoc. Prof. Viktor Medvedev
Laima Paliulionienė
Assoc. Prof. Martynas Sabaliauskas
Prof. Povilas Treigys

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Prof. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. +370 5 2109315

https://doi.org/10.15388/DAMSS.14.2023
ISBN 978-609-07-0985-6 (digital PDF)

Preface

DAMSS-2023 is the 14th International Conference on Data Analysis Methods for Software Systems, held in Druskininkai, Lithuania. Every year at the same venue and time. The exception was in 2020, when the world was gripped by the Covid-19 pandemic and the movement of people was severely restricted. After a year's break, the conference was back on track, and the next conference was successful in achieving its primary goal of lively scientific communication. The conference focuses on live interaction among participants. For better efficiency of communication among participants, most of the presentations are poster presentations. This format has proven to be highly effective. However, we have several oral sections, too. The history of the conference dates back to 2009, when 16 papers were presented. It began as a workshop and has evolved into a well-known conference. The idea of such a workshop originated at the Institute of Mathematics and Informatics, now the Institute of Data Science and Digital Technologies of Vilnius University. The Lithuanian Academy of Sciences and the Lithuanian Computer Society supported this idea, which gained enthusiastic acceptance from both the Lithuanian and international scientific communities. This year's conference features 84 presentations, with 137 registered participants from 11 countries. The conference serves as a gathering point for researchers from six Lithuanian universities, making it the main annual meeting for Lithuanian computer scientists. The primary aim of the conference is to showcase research conducted at Lithuanian and foreign universities in the fields of data science and software engineering. The annual organization of the conference facilitates the rapid exchange of new ideas within the scientific community. Seven IT companies supported the conference this year, indicating the relevance of the conference topics to the business sector. In addition, the conference is supported by the Lithuanian Research Council and the National Science and Technology Council (Taiwan, R. O. C.). The conference covers a wide range of topics, including Applied Mathematics, Artificial Intelligence, Big Data, Bioinformatics, Blockchain Technologies, Business Rules, Software Engineering, Cybersecurity, Data Science, Deep Learning, High-Performance Computing, Data Visualization, Machine Learning, Medical Informatics, Modelling Educational Data, Ontological Engineering, Optimization, Quantum Computing, Signal Processing. This book provides an overview of all presentations from the DAMSS-2023 conference.

PARTNER

International Federation for Information Processing
ifip.org

DAMSS 2023 SUPPORTED BY:

General sponsors

Neurotechnology
neurotechnology.com

Novian
novian.lt

Research Council of Lithuania
lmt.lt

Main sponsors

National Science and Technology Council
nstc.gov.tw

3RTechnology
3rt.lt

Sponsors

Asseco Lithuania
asseco.lt

NetCode
netcode.lt

Vinted
vinted.com

Visoriai Information Technology Park (VITP)
vitp.lt

Literature Review on Data Modelling Patterns and Techniques for Educational Data Assessment

Snieguolė Bagočienė, Anita Juškevičienė

Institute of Data Science and Digital Technologies
Vilnius University

snieguole.bagociene@mif.stud.vu.lt

Educational data modelling can be used to develop theoretical frameworks in educational research, to create predictive models of student outcomes, or to establish guidelines for educational data collection and analysis. This poster presentation provides a review of the literature on data modelling patterns and techniques applied to the assessment of educational data. As education becomes increasingly reliant on data-driven decision making, we explore the scientific work that has contributed to the understanding of this important area. Our study reviews and summarises the key concepts, theories and methodologies in the existing literature. We present a range of data modelling models that have been discussed in academic research, as well as their application in predicting student achievement, identifying learning trends, and adapting educational practices. We also explore the scholarly debate on the importance of data quality assurance and pre-processing in maintaining the accuracy and reliability of educational data models. Following this literature review, we outline best practices and emerging trends in educational data modelling.

Empowering Industrial Control: Soft Computing Paradigms and Practical Applications

Valentina E. Balas

Aurel Vlaicu University of Arad, Romania

balas@drbalas.ro

The lecture delves into the realm of soft computing paradigms for designing intelligent systems. Computational Intelligence, an emerging field, equips us with tools to model and dissect complex systems, encompassing techniques like fuzzy logic, neural networks, genetic algorithms, and more. Presently, fuzzy logic stands as a widely adopted approach to tackle control problems across various applications.

The lecture also showcases how to craft and practically employ intelligent complex systems by integrating deterministic knowledge into the processes and leveraging simulations during the design phase. Through introduced case studies, it vividly illustrates the utility of intelligent control across a diverse array of applications.

Soft computing comprises a suite of methods tailored to handling imprecise information and intricate human cognition. When addressing industrial control challenges, soft computing techniques exhibit notable qualities, including Intelligence, robustness, and cost-effectiveness. This study undertakes the task of providing an extensive overview of soft computing techniques and their roles in industrial control systems. These soft computing methodologies are primarily categorised into fuzzy logic, neural computing, and genetic algorithms.

The study identifies the modern challenges in industrial control systems, such as information acquisition hurdles, complexities in modelling control rules, optimisation difficulties, and the need for robustness. Subsequently, it reviews the advancements in soft computing that have been devised to address these challenges. Moving forward, the study offers a retrospective analysis of practical industrial control applications spanning transportation, intelligent machinery, process industries, and energy engineering.

The presentation underscores how soft computing methods bestow industrial control processes with numerous advantages, highlighting their significant potential applications.

Development of an Integrative Approach to the Assessment of Cognitive Abilities in Patients With Neurodegenerative Diseases of the Central Nervous System

Santa Bartušēvica, Jurģis Šķilters, Solvita Umbraško, Līga Zariņa, Laura Zeļģe, Agnese Anna Pasare, Ardis Platkājis, Jānis Mednieks, Aleksejs Ševčenko, Nauris Zdanovskis, Artūrs Šilovs, Edgars Naudiņš

Laboratory of Perception and Cognitive Systems, University of Latvia
Faculty of Computing, Riga Stradins University, Latvia

santa.bartusevica@lu.lv

This poster presentation describes a research project that is investigating cognitive abilities in patients with neurodegenerative diseases of the central nervous system. The implementer of this project is Riga Stradins University, in cooperation with the Laboratory of Perception and Cognitive Systems at the Faculty of Computing, University of Latvia. The project aims to develop a set of cognitive and perceptual measurement tools that would allow early identification of neurodegenerative diseases based on the comparison of different experimental measurements and tests in a between-group setting where the experimental group consists of patients in clinical conditions (which would also be investigated by standard validated clinical neurological tests prior to perceptual and cognitive measurements) and magnetic resonance imaging (MRI) measurements. Neurodegenerative disorders are diseases of the central neural system that predominantly affect the elderly population and cause severe perceptual, cognitive and motor impairments. They significantly reduce a person's quality of life, often leading to severe and permanent disability. Individual neurodegenerative disorders are heterogeneous in their clinical manifestation and underlying physiology, although they might have overlapping symptoms. Therefore, the precision of diagnosis is essential for determining and monitoring specific treatments. Early diagnosis and treatment can effectively slow down or stop the progression

of these diseases. The main research tasks of this project are (1) to develop a set of non-invasive methods (i.e. cognitive and perceptual tests) for diagnosing early symptoms of neurodegenerative diseases and (2) to understand and identify neural correlations of visual-spatial and verbal processing disorders that could be diagnostically informative in the case of neurodegenerative diseases. These non-invasive methods also include recordings of eye-tracking data with mid-level visual perception (visual organisation) stimuli. The poster presentation will look in more detail at the development and data collection of various digital cognitive tests as well as the use of eye-tracking data in the case of investigating cognitive abilities.

Comparison of Big Data Storage Solutions With the Hybrid Blockchain Architecture

Antanas Bendoraitis, Tomas Blažauskas, Eglė Butkevičiūtė

Department of Software Engineering, Faculty of Informatics
Kaunas University of Technology

egle.butkeviciute@ktu.lt

A massive amount of medical and health data is produced for diagnosis, monitoring, and treatment purposes. The gathered and stored content must be valid, complete, traceable, and immune to modification or deletion. The big data collection from the Internet of Things (IoT) and mobile devices emergence influenced data storage and processing solutions to be more reliable and decentralised. All database management systems are designed to minimise security breach, and it may range from simple password protection to complex user structure design. Wide usage of IoT technologies requires secure data transitions and storage that could be ensured by private blockchain networks that create tamper-resistant records of shared transactions. Currently, there is a lack of comprehensive and specific comparison among widely recognised primary big data storage solutions such as SQL, NoSQL, or blockchain-based approaches. Considering data storage and the frequent issues of data loss that arise from it, the comparison becomes important in security aspects. Furthermore, the other aspects, such as speed, utilised structures, and other possible qualitative measures among alternatives, were also analysed and compared. Also, the presence of such a comparison provides the opportunity to identify more accurately the most suitable alternative when making decisions related to secure data storage. This study proposes a hybrid blockchain architecture that combines traditional SQL and blockchain-based solutions to securely save users' data with tamper-proof resistance. The proposed technique uses an unconventional consensus algorithm for faster agreement and data signing before saving. Primary results seem to be promising since the approximate data-saving time is 3.5 ms where 1 kb of data using 3 nodes is sent to each of the 1000 data packets that contain 1 kb of data.

Similarity Metrics for Cartographic Sentinel-2 Multispectral Imagery Comparison

Algirdas Benetis, Vytautas Valaitis

Vilnius University

algirdas.benetis@gmail.com

Drones localize through GPS signal transmission, but sometimes this is not possible due to interference or noise, and sensors alone are not enough for accurate positioning in the long run. In the era of digitization, many fields, including agriculture or the military industry, use drones for various purposes. Using an orthographic image similarity metric based on triplet neural networks is one way to determine the drone's location. The topic of this study is the calculation and comparison of similarity metrics based on the developed EfficientNet, EfficientNetV2, MobileNet, ResNet and VGG neural network architectures by using different band composites of Sentinel-2. The trained base layers of these networks are used in triplet neural networks. During the experiments of image processing time, distances from the anchor picture and precision metrics, which would help to determine more acceptable architectural configurations and band composites for comparing orthographic images and finding the location of drones, the results of the similarity metrics between band composites were compared with each other. By using combinations of bands, we can extract specific information from an image. E.g., there are combinations of bands that highlight geological, agricultural or vegetation features in an image. The final choice of the triplet neural network model and band combination may depend on various factors and, it is important to emphasize that it is worth considering all the results of the obtained metrics before applying the respective architectures to single cases, to evaluate the importance of each metric and composite in personalized application situations.

Machine Learning Model Evaluation Bias Using A/B Tests in Marketplace Setting

Giedrius Blažys

Vinted

g.blazys@gmail.com

This presentation discusses the assessment of machine learning models in an online marketplace context, specifically at Vinted. It highlights the inadequacy of solely relying on offline performance metrics to gauge model quality, emphasizing the importance of A/B testing in evaluating the real-world impact of machine learning models on business metrics such as conversion rates, transaction values, and service utilization. The presentation also addresses the issue of interference in marketplace experiments, where actions affecting one participant influence outcomes for others. Empirical evidence of such interference is presented, followed by a discussion on the violation of the SUTVA assumption and a review of literature on approaches to manage interference in A/B testing.

Nonparametric Tests on the Adequacy of Models for Count Data

Stefano Bonnini

Department of Economics and Management
University of Ferrara, Italy

bnnsfn@unife.it

In many empirical problems, the response variable of a statistical model is discrete because it represents counting data and it takes non-negative integer values. This is typical of epidemiological studies, in which the goal is predicting the number of cases of a certain disease, as a function of environmental, demographic, social, cultural, geographical, or other factors.

Other frequent applications concern the prediction of the number of committed crimes or occurred road accidents (social research), the number of customers who make certain purchase choices (marketing), the number of defects found on a product (quality control), the number of votes obtained by a candidate or party during elections (politics), the number of extreme weather events that have occurred over a period of time (environmental studies), and many others.

For this type of study, models for count data, based on the Generalized Linear Models (GLM) approach, are typically applied. The most commonly used are the log-linear models, in particular, the "Poisson" and "Negative Binomial" models. The tests of hypothesis on the validity of such models are based on the main assumption that the response follows a specific probability distribution and they are not performant when this assumption does not hold.

In this work, we propose a nonparametric test on the validity of the model, which is based on the permutation approach. The good performance of the test, in terms of power behavior, compared to those of the classic parametric tests, is proved through a Monte Carlo simulation study. The results of its application to a survey regarding the effect of public policy interventions on the adoption of 4.0 technologies by Italian enterprises are also presented.

This research is part of the project entitled “Public policies, 4.0 technologies and enterprise performance. Empirical analyses on a representative sample of manufacturing enterprises of northern Italy”, funded by the University of Ferrara for the period 2022-24, through the Departmental Research Incentive Fund – FIRD 2022.

What Is a Concept Drift, and Does It Affect Machine Learning Performance?

Dalia Breskuvienė, Gintautas Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University

dalia.breskuviene@mif.vu.lt

Traditional machine learning algorithms are built expecting that data for training and testing has roughly the same proportion of classes. Additionally, data and target distribution are expected to not change over time. However, this is not the case in many real-world situations, such as fraud detection, medical diagnosis, natural disaster prediction, or manufacturing quality control.

A phenomenon where the statistical properties of the target variable or data distribution change over time is called Concept Drift, meaning that models trained on historical data become less effective. We aim to comprehensively explore the concept drift phenomenon, its manifestations, and its impact on machine learning performance. This study describes concept drift and discusses its significance in real-world fraud detection systems. It delves into the causes and drivers of concept drift, including evolving user preferences, environmental changes, and intrinsic data dynamics.

We provide a comprehensive overview of concept drift, highlighting its importance and challenges in machine learning. It emphasizes the need for robust and adaptive models to address concept drift in dynamic systems and discusses the implications of this phenomenon in the broader field of artificial intelligence. Understanding and effectively managing concept drift is crucial for successfully deploying machine learning systems in evolving and complex domains. Our review focuses on the state-of-the-art techniques and approaches to detect and adapt to concept drift, focusing on incremental, transfer, and diversity ensemble learning.

To quantify the effect of concept drift on machine learning performance, we present experimental results based on benchmark datasets and real-world case studies in the literature. These results reveal how concept drift can degrade model performance, increase false positives, and lead to unexpected consequences in dynamic environments.

An Overview of Holographic Data Representations

Alfred Bruckstein

Technion Israel Institute of Technology, Israel

freddy@cs.technion.ac.il

Holographic representations of data encode information in packets of equal importance that enable progressive recovery. The quality of recovered data improves as more and more packets become available. This progressive recovery of the information is independent of the order in which packets become available. Such representations are ideally suited for distributed storage and for the transmission of data packets over networks with unpredictable delays and or erasures.

Insider Threat Detection: A New Keystroke Dynamics-Based Approach to User Authentication in Critical Infrastructure

Arnoldas Budžys, Olga Kurasova, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

arnoldas.budzys@mif.stud.vu.lt

In today's evolving digital landscape, challenges such as unauthorised intrusions, cyber security breaches, and data compromise are threatening national defence, critical infrastructure, and economic sustainability. Robust authentication mechanisms are essential to address these vulnerabilities. Researchers are exploring the challenging problem of protecting critical infrastructure from insider threats, especially in the context of their increased levels of access and trust. To address this challenging problem, we present a deep learning-based methodology for user authentication. Our methodology is based on transforming keystroke time series data generated from passwords (numerical data) into images to increase the efficiency of intrusion detection and the accuracy of user verification. We present the GAFMAT (GAbor Filter MAtrix Transformation) method, a new approach for transforming numerical password data into a visual format. The Siamese neural network architecture with triplet loss function (or triplet network) is used to detect abnormal or unauthorised login entries. The authentication process based on this architecture compares the features of the password entered by the user and, using the proposed method, transformed into an image with previously known and transformed records, evaluating the authenticity against the reference transformed passwords stored in the database. We have proven the effectiveness of the GAFMAT method by using publicly available datasets and transforming them into visual representations. The experiments resulted in competitive and, in many cases, better EER (equal error rate) values compared to existing machine learning techniques. The robustness of GAFMAT to a range of passwords of different lengths and complexity confirms its potential as a reliable method for biometric authentication based on keystroke dynamics.

Classification of Satellite Images to Create a Map of the Settlement Area

Laurynas Buinauskas[1], Aušra Gadeikytė[2], Eglė Butkevičiūtė[1]

[1] Department of Software Engineering, Faculty of Informatics
Kaunas University of Technology
[2] Department of Applied Informatics, Faculty of Informatics
Kaunas University of Technology

ausra.gadeikyte@ktu.lt

The growth of deep learning technologies has transformed image classification, especially by improving its ability to analyze satellite imagery for accurate identification of urban structures. However, satellite imagery classification of urban structures is often limited due to resolutions, building shape variances, overlapping, the complexity of the background, etc. This study aims to use advanced deep learning algorithms to automate the classification of satellite-taken land images, in order to facilitate the creation of detailed urban area maps. The study was conducted using object detection algorithms like AlexNet, VGGNet, InceptionNet, and ResNet. The main task was to achieve high model accuracy and efficiency. For this reason, the methodical data preparation strategy includes resizing and processing the original images. It should be noted, that in such a manner the dimensions of the obtained images are divisible by two, while also ensuring the representation of the original imagery without losing too much information. Concurrently, the generation of binary masks delineating building footprints is essential for data preparation. Various pre-processing strategies were explored to augment the model's learning efficacy from the data. Classification of satellite images was done using the TensorFlow machine-learning framework. Metrics such as Intersection Over Union (IOU) and Pixel Accuracy were evaluated to ensure the accuracy of building segmentations in satellite images. The proposed classification system demonstrates the potential to efficiently identify buildings from other objects in each area that could be used for more precise urban planning and infrastructure development strategies. The presented system is capable of generating accurate urban maps and identifying unauthorized constructions, thereby contributing to enhanced infrastructure planning.

Adaptive Mapping of Cybersecurity Competence Assessment Methods

Karina Čiurlienė

Vilnius Gediminas Technical University

karina.ciurliene@vilniustech.lt

The growing number of security threats requires more sophisticated ways to protect IT infrastructure and systems. However, technical measures do not provide adequate protection and must be combined with sociological solutions and skills. Knowledge and competence in the field of cybersecurity are one of key factors that allow increasing cybersecurity assurance in an organization, therefore, education of regular users as well as security professionals must be of high priority. The education process has its lifecycle and includes knowledge and competence assessment. When the cybersecurity competence assessment is finished, then the selection of corresponding learning material content can be made. Such a way of education increases its quality and reduces the time used for learning. On the other hand, the knowledge and competence demonstrated during the assessment depends on the type of assessment method as well as social-psychological aspects such as emotional environment, stress, motivation, etc. Gathered data about the dominant characteristics and risks of cybersecurity specialists or regular users makes it easier to select the corresponding competence assessment method. The goal of this research was to make the analysis of cybersecurity competence assessment methods based on data collected using questionaries, interviews, and observations. Educational and social-psychological aspects are included in the datasets. Also, the adaptive mapping of cybersecurity competence assessment methods was proposed based on the following data: user profile, experience, technical background, stress, emotional robustness, and motivation. Moreover, Bloom's taxonomy was used for mapping cybersecurity competence assessment methods. It helps to understand the relation between the competence model and assessment methods as well as complements cybersecurity training programs.

Obfuscation and Evasion Techniques for Red Team Assessments

Juozas Dautartas, Arnoldas Budžys, Viktor Medvedev

Institute of Data Science and Digital Technologies
Vilnius University

juozas.dautartas@mif.stud.vu.lt

In today's increasingly complex digital environment, businesses, governmental institutions, and ordinary citizens can become a target of cyber criminals. Therefore, measures like advanced Anti-viruses, Endpoint Detection and Response systems, and Extended Detection and Response systems are becoming more and more critical in everyday life as successful cyber-attacks can cause severe damage (e.g., Not-Petya attack in 2017). That's why large organizations have their cyber defense specialists working around the clock in what as part of so-called Blue teams. In many cases, these specialists protect critical infrastructure such as banking sectors, power plants, governmental infrastructure, or businesses in general. Moreover, these Blue teams usually rely heavily on previously mentioned security tools and the telemetry that these tools gather. Therefore, it became a common practice to hire ethical hackers who try to breach and test Blue team's effectiveness. Additionally, report these weak points to security teams before any cyber criminals exploit these holes.

To simulate real-world attacks, Red teams usually use open-source or custom tools to achieve their goals. Since modern defense tools commonly use advanced machine learning algorithms to detect malicious activity, strong malware obfuscation and evasion techniques are particularly important for realistic adversary emulation. In this work, a concept of „ethical malware" obfuscation will be introduced to validate and strengthen existing security defences. Using machine learning techniques, our approach combines generative adversarial networks (GANs) and Siamese neural network capabilities to create, validate, and identify obfuscated malware. The essence of this ethical malware is that it evades detection by traditional defenses. It also intends to work on specialized

malware feature extraction and methods for transforming non-image data into visual form (e.g., GAFMAT method) for training convolutional neural networks and generating malware using GANs.

The effectiveness of these methods could be tested in national and NATO cyber security exercises such as Amber Mist and Locked Shields. Overall, this research is intended to contribute to more resilient and adaptable cyber security as well as train high-level professionals to seek out emerging cyber threats.

Comparative Analysis of Eleven Dynamic Artificial Neural Networks

Martynas Dumpis, Dalius Navakauskas

Vilnius Gediminas Technical University

martynas.dumpis@vilniustech.lt

With the increasing complexities in sequential data handling and interpretation, dynamic artificial neural networks have gained significant prominence. The structures incorporating time delays in their synapses have shown potential in various applications ranging from natural language processing to financial forecasting. However, a comprehensive comparative study explaining the differences, advantages, and potential drawbacks of dynamic neural networks remains a gap in the literature.

This research provides a detailed comparison of 11 dynamic neural networks known for their time-dependent characteristics: Time Delay Neural Network, Time Derivative Neural Network, Bi-directional Neural Network, Finite Impulse Response Multi-Layer Perceptron, Simplified Finite Impulse Response Multi-Layer Perceptron, Infinite Impulse Response Neural Network, Gama Memory Neural Network, Lattice Ladder Neural Network, Recurrent Neural Network, Long Short-Term Memory, and Gated Recurrent Unit.

To comprehensively understand the performance and properties of each dynamic neural network, this comparison is based on several pivotal metrics. First of all, the accuracy of each model is assessed to understand its predictive correctness. The magnitude of the model's error over epochs, known as loss, provides further insights into its learning progression. Another significant aspect we observe is the training time which reveals how efficiently a model can adapt and learn using gradient descent. The size of the model, determined by the number of parameters, influences deployment and scalability considerations, making it another crucial factor in our analysis. Furthermore, the robustness of each dynamic neural network is tested against noisy or adversarial conditions, while their tendencies towards overfitting give us a clear picture of their generalization capabilities versus the risk of memorizing training data. Lastly, scalability is probed by examining how each dynamic neural network's performance evolves as the complexities or sizes of datasets increase.

Development of the NO-GAP Educational Analytics Tool: Evaluating Disparities in Academic Achievements Among Lithuanian Students

Rasa Erentaitė[1], Rimantas Vosylis[1], Daiva Sevalneva[1], Eglė Melnikė[1], Vaidas Morkevičius[1], Giedrius Žvaliauskas[1], Berita Simonaitienė[1], Monika Zdanavičiūtė[2,3], Anton Volčok[2,3], Tomas Krilavičius[2,3]

[1] Kaunas University of Technology
[2] Vytautas Magnus University
[3] Centre for Applied Research and Development

monika.zdanaviciute@card-ai.eu

Development of the NO-GAP Educational Analytics Tool: Evaluating Disparities in Academic Achievements among Lithuanian Students" One of the primary challenges faced by modern educational systems is the disparity in students' academic achievements. This disparity manifests as long-term significant differences among student groups and academic institutions. Simply understanding the factors influencing students' academic achievements at a population level is insufficient. Therefore, it is crucial to investigate how various achievement factors manifest differently among specific student groups. We have developed the student achievement analysis tool, named "NO-GAP," designed to facilitate in-depth analyses of student data by school personnel. This tool enables the analysis of the achievements of students and the disparities of achievements, as well as the reasons for them. Furthermore, it allows for the analysis of the relations between different achievement indicators according to the school contexts. The methodological framework for data analysis and visualization was created using a data set consisting of 773 schools and 25,976 students in Lithuania, which includes data on the academic achievements of students from 2017 to 2022.

Visual Object Recognition – Traditional Methods Along With Deep Learning Approaches

Jan Flusser

Institute of Information Theory and Automation, Czech Republic

flusser@utia.cas.cz

The talk falls into the area of visual Artificial Intelligence (AI), particularly to image recognition by deep networks. In AI applications such as surveillance systems, autonomous robots, unmanned vehicles, drones, etc., cameras and other visual sensors form the "eyes" of the system while image recognition algorithms substitute the visual cortex of the brain. The key requirement is a continuous (possibly real-time) analysis of the visual field and, in that way, preparing the basis for decision and next action planning. The visual analysis may comprise scene segmentation, detection of objects and persons of interest, recognition of their identity and their behaviour, and even prediction of their next actions. In this talk, we focus on the recognition part, where the image/object is classified as a member of one of the pre-defined classes. Current convolutional networks work with inefficient pixel-wise image representation, which does not provide almost any invariance. This leads to the use of very large training sets and to massive augmentation. We propose to decompose intra-class variances into two degradation operators where one of them (image rotation, scaling, blurring, etc.) can be mathematically modelled by a superposition integral with a transformation of the coordinates. We further propose to design hybrid network architectures that use both pixel-level and newly developed high-level invariant image representations such that the high-level representation eliminates the influence of modelable degradations. This leads to a substantial reduction of the training data without sacrificing the recognition rate. The hybrid architectures could define a new standard in image-oriented networks.

Evolutionary Model for Non-Coding Nucleotide Sequences

Melita Frolovaitė[1], Elise Manon Marie Lebon[2], Tomas Ruzgas[1]

[1] Kaunas University of Technology, Lithuania
[2] University of Angers, France

tomas.ruzgas@ktu.lt

During the past decades, due to the fundamental discoveries in the field of molecular biology, it has become a central subject of biology sciences. Previous focus of attention has been shifted from the identification of one specific gene to greater opportunities that have become possible by the sequencing of complete genomes. That, in turn, opened the door to the technologies of the so-called post-genomic era. They are often based on the computer analysis of the entire genome, i.e. on bioinformatics.

The numbers of nucleotides and amino acids in such databases of nucleotide sequences as *GenBank*, *DDBJ* or *EMBL* have been continuously increasing and have become enormous. With such extensive and continuously supplemented data amounts available, recognition of biological signals in an individual nucleotide sequences or the whole DNA, as well as their determination of their function have become a complicated task and a relevant problem of bioinformatics.

Until now, any randomised sequence of nucleotides or amino acids was considered to be a noncoding nucleotide or amino acid sequence. The work offers and substantiates the opinion that prior good knowing of biological-"genetic noise" is necessary to detect a biological signal in DNA sequences. There occurs a need for definition and accurate formulation of the notion of "genetic noise".

The statistical analysis carried out in the work reveals that the major part of even non-coding nucleotide sequences are not of the first order Markov chain, which is serious grounds for having doubts about the available models of nucleotide sequences, assumptions of their existence and adequacy of their application. This means that, for example, a comparison of real sequences with ones generated according to such

models is not a reliable tool in the search either a biological signal or a biological function of a specific nucleotide (or amino acid) sequence. The same holds regarding the accuracy of phylogenetic trees reconstructed by means of these models. As an alternative for the existing models, a mathematical definition of noncoding nucleotide sequence or, in other words, of "genetic noise", has been formulated and its model has been proposed.

The theory of discrete Markov fields is used to define a noncoding nucleotide sequence and to formulate its properties. The model of the noncoding sequence is verified by computer simulation of nucleotide sequence evolution. To analyse DNA sequence structure and nucleotide dependence, correlation and R/S (rescaled range) analysis are used. To verify Markovity of a nucleotide sequence, loglinear and generalised logit models are applied and appropriate hypotheses are verified on their basis.

Evaluating End-to-End and Multi-Stage 3D Model Generation With 3D Morphable Models: A Comparative Study on Eyeglasses Frames

Ervinas Gisleris, Artūras Serackis

Department of Electronic Systems
Vilnius Gediminas Technical University (VILNIUS TECH)

arturas.serackis@vilniustech.lt

In the domain of product visualization, the ability to generate precise 3D models from 2D images is pivotal for enhancing user experience and engagement. Our investigation conducts an exhaustive comparative analysis of two distinct methodologies in generating 3D models of eyeglasses frames from a pair of product images: the end-to-end 3D model generation and a multi-stage approach utilizing 3D Morphable Models (3DMM). The end-to-end generation approach leverages deep learning algorithms to directly map input images into 3D models, ensuring an uninterrupted and expedited transformation. On the other hand, the multi-stage approach begins with the extraction of pertinent features from the 2D images, followed by segmentation and the application of 3DMM for shape reconstruction. The 3DMM technique is adept at capturing nuanced geometric variations, enabling a more tailored and accurate modeling of eyeglasses frames. Our analysis closely evaluates the two methodologies based on crucial metrics such as accuracy, computational efficiency, and fidelity to the original product images. By concentrating on eyeglasses frames, which often present intricate and subtle design features, we aim to ascertain the effectiveness of each method in accurately capturing and reconstructing these complex structures.

The findings of this comparative study offer valuable insights into the strengths and limitations of each approach, providing guidance for businesses and researchers seeking to optimize 3D model generation for e-commerce applications. The results also pave the way for further improvements in ensuring accurate, efficient, and user-friendly solutions for product visualization and virtual try-on applications.

On Computational Challenges in Value Iteration for Inventory Control

Eligius M. T. Hendrix

University of Malaga, Spain

eligius@uma.es

Inventory control has always been an interesting subject for mathematic and stochastic analysis, simulation and computation. The concept enhances rules that tell us how much to order (replenish) from which product in which situation of the inventory level. On a personal level, you should decide how much beer to store in your fridge and when to replenish it by going to the shop. Challenges are getting bigger when we have perishable products like lettuce or strawberries in them as we have to take care of their age and consume them before a due date. On the level of retail, as one-third of world food production is disposed of every year, there is a renewed interest in deriving adequate inventory control policies for perishable products. Simulation-based optimisation and Stochastic Programming approaches may be used to derive optimal control rules. A computational challenge appears when in the order rules we would like to take the age distribution of items in stock into account. Dynamic programming is an elegant tool to derive the best rule. Specifically, we will sketch the concept of Value Iteration for small cases. It is based on the so-called Bellman optimality criterion for dynamic programming. Under certain conditions, the value iteration may lead to the optimal rule. In a practical sense, dynamic programming is confronted with the so-called curse of dimensionality. This means that if more factors are taken into account in the state space, it is hard in an exponential sense to come to an optimal solution. Here, careful bounding is relevant. The target of the presentation is to sketch the idea with very small instances and gradually illustrate what happens if, for instance, the age of products is taken into account, the lead-time of delivery or the amount sold, which can be returned to the shop. We use several small instances to showcase the underlying characteristics of the corresponding optimisation problem.

A Comparative Study of Mathematical Models for Aircraft Deconfliction Problem

António Iglesias, Tatiana Tchemisova

University of Aveiro, Portugal

tatiana@ua.pt

In the field of air traffic management, ensuring a safe distance between flying aircraft while optimising specific objectives is a critical challenge known as the Aircraft Deconflict Problem (ADP). Traditionally, human Air Traffic Controllers (TAC) handle this task by adjusting aircraft altitudes, trajectories, directions, or speeds to resolve potential conflicts in restricted airspace. However, there is a growing interest in introducing automation to aircraft deconfliction, including the emerging concept of urban air mobility. One intriguing approach is the concept of subliminal speed control, which involves subtly adjusting the aircraft's speed in a way that might not always be readily detectable by TACs. This approach is aimed to efficiently reduce the number of conflicts in the airspace before direct control. In this research, our main focus lies in obtaining numerical results using Semi-infinite Programming (SIP) and Nonlinear Programming (NLP). To solve this SIP and NLP models, we will use Python and MATLAB software, particularly the SciPy and Optimization Toolbox packages, and then comparing their final results.

Automated Verification of Railway Signalling Data

Alexei Iliasov, Dominic Taylor,
Linas Laibinis, Alexander Romanovsky

Institute of Computer Science
Vilnius University

linas.laibinis@mif.vu.lt

SafeCap is a modern toolkit for modelling, simulation and formal verification of railway networks. This presentation discusses the use of SafeCap for formal analysis and automated, scalable safety verification of solid-state interlocking (SSI) programs – a technology at the heart of many railway signalling solutions around the world. The main driving force behind SafeCap development was to make it easy for signalling engineers to use the technology and thus to ensure its smooth industrial deployment. The unique qualities and the novelty of SafeCap are in making the use of formal notations and proofs fully transparent for the engineers. The presentation explains the formal foundations of the proposed method, its tool support, and its successful application by railway companies in developing industrial signalling projects.

The Use of Smartphone Data in Symptom Identification for Patients With Cancer

Gabrielė Jenciūtė[1,2], Gabrielė Kasputytė[1,2], Nerijus Šakinis[1,2], Paulius Savickas[1,2], Inesa Bunevičienė[1], Erika Korobeinikova[3], Domas Vaitiekus[3], Arturas Inčiūra[3], Laimonas Jaruševičius[3], Ričardas Krikštolaitis[1], Tomas Krilavičius[1,2], Elona Juozaitytė[3], Adomas Bunevičius[3]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development
[3] Lithuanian University of Health Sciences

gabriele.jenciute@vdu.lt

Cancer patients experience physical and psychological challenges resulting from either the disease or the toxic effects of cancer treatments. Early detection and prediction of symptom onset or exacerbation can significantly enhance patient outcomes by preventing disease progression and maintaining their quality of life, while remote monitoring has emerged as a crucial tool in comprehensive care of cancer patients. In this study we explore the possibility of using passively generated sensors data in patients' behavior analysis and questionnaire data in symptoms assessment. Passively, and continuously generated data-streams from smartphone sensors were used to describe the activity and sociability of 108 patients with different type of cancer. Symptom severity was evaluated using the European Organization for Research and Treatment of Cancer Core Quality of Life questionnaire. Time series models: Autoregressive Integrated Moving Average, Holt Winter's method, TBATS, Long Short-Term Memory Neural Network and General Regression Neural Network were used to create a methodology for symptom identification. The analysis revealed that three symptoms – depression, fatigue, and vomiting correlate to passively collected smartphone sensor data. This project has received funding from European Regional Development Fund (project No. 01.2.2-LMT-K-718-05-0085) under grant agreement with the Research Council of Lithuania (LMTLT).

Modelling the Evolution of Relational Event Data

Rūta Juozaitienė[1,2], Ernst C. Wit[3]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development
[3] Università della Svizzera Italiana

ruta.juozaitiene@vdu.lt

Relational event networks naturally emerge from social interaction: interactions can be considered as time-stamped edges between vertices of social actors. By studying the factors that influence the interaction inter-arrival times, we can get insight into the drivers of interaction dynamics. Relational event models provide a succinct way to analyze a broad family of interactional patterns and their influence on network dynamics. These models can easily incorporate a wide range of mechanisms involving both endogenous and exogenous network effects. Traditionally, quantitative models have assumed that endogenous network mechanisms, such as reciprocity or triadic effects, remain constant over time. However, a number of studies argued that these effects are likely to be dynamic over time. For instance, reciprocity often exhibits a strong tendency to decay over time. To properly capture the evolving nature of the network effects, we propose employing a relational event model that incorporates time-varying effects in terms of smooth functions by offering a computational edge over the models in the literature.

Towards Effective and Efficient AI-Assisted/Enabled Decision Aid and Support Systems: A Challenge for the Human-AI Collaboration

Janusz Kacprzyk

Systems Research Institute
Polish Academy of Sciences, Poland

Janusz.Kacprzyk@ibspan.waw.pl

Artificial intelligence (AI), with its presumably most relevant subfield of machine learning, is often considered to be the next Industrial Revolution and is taking by storm virtually all areas of science and technology. Since decision-making is the most frequent act in human activities and its role in all aspects and activities in our life is crucial, it has become obvious that AI will also be a decisive factor in the quest for making better decisions. We consider decision-making processes in complex environments, with many stakeholders, criteria, dynamics, etc., and advocate the use of smart decision aid and decision support systems, and in particular, advocate the use of AI as their driving force. First, we present new directions in AI-assisted decision-making, showing the role of data-driven approaches to deal with both big data and small data type problems, notably aimed at supporting decision-making. We advocate the use of a labour division by including domain and tool specialists, which can be implemented via a decision aid architecture and show problems with such a multi-stakeholder situation. Then, we assume a more „democratic" approach of an AI-enabled DSS (decision support system), which does not require as much expertise from the stakeholders as the decision aid. We also emphasize the data-driven approaches and a wide use of machine learning to derive all kinds of the users' characteristic features. We also consider the context awareness and intention awareness aspects. Emphasis is on the broadly perceived human-AI collaboration exemplified by a proper account for the basic difference in the number crunching power of the computers and human capabilities of solving

more complex problems, human deficiencies in the sense of cognitive biases, various dilemmas related to fast and slow decision making, etc. An important aspect will be devoted to a proper selection of problems or their parts that should be solved by either the human or a computer system. Some future challenges and directions will be briefly mentioned.

Assessment Functions Accounting for Attribute Balance

Ignacy Kaliszewski

Systems Research Institute of the Polish Academy of Sciences, Poland

ignacy.kaliszewski@ibspan.waw.pl

We investigate assessment functions, i.e., functions that aggregate numerical attributes into single numbers. Varying by application domains, multiattribute assessment functions are termed "value functions", "utility functions", "production functions", or "scalarizing functions". All assessment functions in the current use share the same limitation; they do not account for attribute proportionality; thus, the issue of *attribute balance* escapes them. We present functions that allow us to account for attribute balance. However, these functions are at odds with the well-established paradigm of Pareto efficiency. The relevance of those functions to rankings is discussed.

Aerial Image Similarity Estimation Using Cloud Removal Methods

Dominykas Kaminskas, Vytautas Valaitis

Vilnius University

dominykas.kaminskas@mif.stud.vu.lt

Unmanned aerial vehicles (UAVs) are used in farming, traffic control, police operations. One of the challenges UAVs faces is the loss of GPS signal. To combat the problem, aerial vehicles use built-In sensors and cameras to help navigate and calculate flight trajectory. This article (research) discusses a map-based approach for aerial vehicle localization: images taken by onboard cameras during flight are compared to an aerial map to find similarities between them. However, the accuracy of a map-based approach decreases during cloudy weather conditions. Cloud coverage is considered a significant loss of information. It can become an obstacle when comparing satellite imagery and lead to deviation in flight trajectory. Removing clouds using neural networks and generative image inpainting algorithms can increase the amount of information found in aerial images. Modern techniques require additional data such as multispectral satellite imagery or cloud-free pictures taken over different time intervals to fill in cloudy image areas. This article introduces a method capable of cloud detection and removal using only RGB bands. Various experiments based on convolutional neural networks and the triplet loss function were conducted to prove the effectiveness of cloud removal methods in aerial image similarity tasks. After testing numerous network configurations results demonstrate that aerial images with cloud removal algorithm applied to them outperform original cloudy images.

Large Language Models for Multilingual and Cross-Lingual Chatbot Communication

Jurgita Kapočiūtė-Dzikienė

UAB Tilde IT, Vytautas Magnus University

jurgita.k.dz@gmail.com

This summary presents chatbot research conducted within three R&D projects: the national 'BotCloud', and two Horizon projects, 'COMPRISE' and 'StairwAI', spanning from 2018 to 2023. Our objectives were to: 1) develop NLU modules for accurate user query understanding; 2) enable NLU modules to comprehend multiple languages; 3) create high-performance NLU modules; 4) address challenges related to limited customer training data, often with few instances. Language coverage in these projects: BotCloud – LT, LV, EE, EN, RU; COMPRISE – EN, GE, FR, LT, LV, PT; StairwAI – EN, GE, ES, FR, IT, LV. Initially, the focus was on monolingual NLU models, but it shifted a few years ago to meet the demand for multilingual communication. Despite having a limited number of intents and the ability to machine-translate chatbot responses, the challenge arises from the diverse ways users ask questions in different languages. To address this, we seek multilingual and cross-lingual solutions without the necessity for training data in each language. In our research, we conducted various comparative experiments using different training and testing strategies: monolingual (training and testing in the same target language), cross-lingual (training in one language, testing in another), combined (training in English and the target language, testing in the target language), and multilingual (training in several languages, testing in the target language). We explored word and sentence embeddings with classification, semantic similarity-based, and generative models based on large language models like Google's BERT (mBERT, LaBSE) and OpenAI's GPT (ADA, Davinci). Our experiments involved diverse model architectures and hyper-parameter optimisation, ranging from small-scale training to unfreezing 0.5 billion parameters. The results of our experiments emphasise the value of sentence embedding models with optimised classifiers or similarity-based techniques. The most accurate NLU modules are integrated into chatbot systems used by Tilde's customers.

Loop Decomposition-Based Integer Programming Model for RNA Secondary Structure Prediction

Olga Karelkina

Systems Research Institute of the Polish Academy of Sciences, Poland

karelkin@ibspan.waw.pl

Nucleic acid secondary structure prediction is an essential problem in the domain of molecular biology since structure can provide insight into structure-function relationship and tertiary structure. It is well known, that RNA molecules in their natural environments tend to fold into their minimum free energy secondary structure.

To predict the energetically optimal secondary structure, an integer linear programming model is suggested based on loop decomposition. The secondary structure of any given RNA sequence can be analytically decomposed uniquely into several characteristic substructures: stem, hairpin, internal, bulge, and multi-branch loops. The free energy of loops that potentially can be formed is measured based on Turner's nearest-neighbour rules and energy parameters. For any given RNA sequence, folding total free energy is calculated as a sum of energies of individual structural elements it is decomposed into. The conformation space of biologically feasible structures is defined in the model by a set of constraints.

Computational experiments conducted on small and middle-size instances of RNA sequence demonstrate that an integer programming model based on loop decomposition can generate a set of alternative secondary structures with energy close to or equal to reference structures' energy values.

Minimum-Sum-Of-Squares Clustering With (Net) Constraints for Cluster-Centres

Mindaugas Kepalas, Julius Žilinskas

Institute of Data Science and Digital Technologies
Vilnius University

mindaugas.kepalas@mif.stud.vu.lt

We present our research results on the problem indicated in the title. Minimum-sum-of-squares clustering is a famous data-science problem for which locally optimal solutions can be found by running the famous k-means algorithm. However, if one seeks to find the global (the best possible) solution of the problem, things become really challenging: this task in scientific literature is proven to be NP-complete. In our research, we modify the famous min-sum-clustering problem by introducing constraints for the placement of cluster centres: these must be placed in a subset of the space (as opposed to the original problem, where the centres can be placed anywhere in the space). In 2-dimensions, possible restrictions might be, for example, that the centres must be placed on a road (e.g., on a net), or cannot be placed in certain regions (e.g., lakes or forests). We report the results of our attempts to find the global solution to the presented problem.

Comparative Analysis of Homogeneity Tests for Censored Samples Under Crossing of Survival Functions

Gintarė Klimantavičiūtė, Rūta Levulienė

Institute of Applied Mathematics
Vilnius University

gintare.klimantaviciute@mif.stud.vu.lt

Survival analysis is a statistical method used to analyze data where the time to a specific event is studied. In this work, a power simulation study was performed using modeling to compare homogeneity criteria for censored samples, when the survival functions may intersect. The following criteria were examined: log-rank ([2]), a two-stage procedure (TSPV), proposed by Qiu and Sheng (see [3]), modified log-rank (MLR), and modified informative criterion (MS) proposed by Bagdonavičius et al. ([1]). Modeling was conducted with various sample sizes and different distribution functions, covering various scenarios when survival functions do not intersect, intersect at the beginning, in the middle, and at the end of the time interval. To explore at what sample size these criteria can be reliably applied, a significance level computation was performed. The results of the analysis show that the power of the criteria depended on the specific characteristics of the simulated data. However, it was found that MLR and TSPV criteria performed best in various scenarios. The results provide researchers with recommendations on which statistical method to use when comparing survival curves with censored samples. Real data analysis was also conducted to illustrate the application of these criteria.

References

[1] Bagdonavičius, V. B., Levuliene, R. J., Nikulin, M. S., Zdorova-Cheminade, O. Tests for Equality of Survival Distributions Against Non-Location Alternatives. Lifetime Data Analysis vol. 10, 2004, p. 445–46.

[2] Mantel, N. Ranking Procedures for Arbitrarily Restricted Observation. Biometrics, vol. 23, no. 1, Mar. 1967, p. 65-78.

[3] Peihua, Q., Sheng, J. A Two-Stage Procedure for Comparing Hazard Rate Functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 70, no. 1, Nov. 2007, pp. 191-208.

Constructing Production Planning Tasks: A Case Study of "Teltonika"

Rima Kriauzienė, Andrej Bugajev

Vilnius Gediminas Technical University

kriauziene@gmail.com

Production planning is a pivotal domain that determines the efficiency and competitiveness of modern enterprises in today's market. The aim of this presentation is to discuss how mathematical modelling can be applied in the construction of production planning tasks, using the "Teltonika" case as an example. In the main formulation of the production planning task, critical parameters such as the number of orders, quantity of products, number of workers in each stage, and the number of stages were identified. These data provide a profound insight into the timelines of the production process, product delivery deadlines, and distribution of workers. The detailed formulation of the task incorporated mathematical nuances, like binary variables and time parameters. Mathematical modelling techniques were employed to describe the production process in the most accurate manner, facilitating a better understanding and analysis of production workflows. The essence of the presentation is to highlight how a detailed task formulation is vitally important for the proper understanding and analysis of production processes. Emphasis is placed on how mathematical structures and methods enable a precise and comprehensive description of the production process.

Challenges in the Next Generation (Oxford Nanopore) Direct RNA Sequencing (dRNA-seq) Data Processing: Coping With Normalisation of Gene Expression Levels by Principal Component Analysis

Algimantas Kriščiukaitis, Robertas Petrolis, Rugilė Dragunaitė, Rytis Stakaitis, Daina Skiriutė

Lithuanian University of Health Sciences

algimantas.krisciukaitis@lsmu.lt

Next generation sequencing offers new possibilities of genetic sequencing at comparatively low cost and make it widely applicable in modern medical diagnostics based on molecular markers of numerous diseases. However, processing and evaluation of sequencing data is challenging due to some inescapable physical factors causing certain bias in the analyzed data. Technical differences ("batch effects") as well as differences in sample processing (up to RNA extraction), and RNA-seq library preparation between the data sets may significantly affect the ability to draw generalizable conclusions from such studies. Gene expression level evaluation across several separately processed samples is the task where the results could be biased by technically caused differences in the estimates. The key to normalization of gene expression levels estimated across different samples lays in so called housekeeping genes, which are required for the maintenance of basal functions that are essential for the existence of a cell. Thus, they are expected to be equally expressed in all cells. Nowadays more than 3000 genes in human genome are defined as housekeeping ones and expression values of at least 10 out of them are recommended to be used as reference values for normalization across the investigated samples. However, there is quite big variety in absolute expression levels even between these housekeeping genes and there is no single recipe how the normalization should be done.

We propose Principal Component Analysis approach to search for gene expression level based biomarkers analyzing Oxford Nano-pore

sequencing data. The principal components, which reflect maximal difference between pathological and control samples, are checked whether they contain housekeeping genes and to what extent. Ideally, the principal component not containing any housekeeping gene should be considered for future evaluation finding the most important genes, the expression of which could be used as biomarker of investigated pathology.

Our experiments were carried out on sequenced data from high, low-grade glioma and healthy control cells. The expression levels of housekeeping genes were present in several first principal components, but usually one component contained them at significantly lesser extent at the same time showing significant difference in pathological vs control samples. The most important expression levels in that component were in the list of expected candidate genes to be a molecular biomarker of investigated disease, selected by other studies.

Balancing Techniques Influence in Financial Distress Detection

Dovilė Kuizinienė, Tomas Krilavičius

Vytautas Magnus Universityy

dovile.kuiziniene@vdu.lt

Financial distress is a condition of an enterprise that has difficulties fulfilling its financial obligations. Financial difficulties can be expressed not only through the bankruptcy of the enterprise but also through the decrease in net income, earnings, ROA, total assets, etc. Bankruptcy is the last stage of financial distress; therefore we want to include more accurate class identifications. Financial distress is indicated through the institution's, debt, employees, and financial statement (net income and equity) indications in this research. This condition expansion allowed class events to increase from ~3% to ~10%. However, a large imbalance between classes remains. Thus, to effectively train the machine learning algorithm, it is necessary to apply class-balancing techniques. To analyze it in Financial Distress detection, we have used ten different balancing techniques: five oversampling (Adasyn, GAN, Smote, SmoteNC, Rose), three undersampling (Random, Nearmiss, K-mean), and two both techniques sampling mix (SmoteENN and SmoteTL). The date used in this researched has been provided by LTD "Baltfakta". It contains 64,428 active small, medium size enterprises in Lithuania, covering the period from 01–01–2015 to 30–12–2022. There are 183,944 unique records in the final data set, which is divided into training and testing sets. Class identifiers covering the years 2018 through 2021 are included in the training set, whereas the testing data set covers the most recent era, meaning the class variable is based on the year 2022. As a result, a data set is split into testing and training sets at a ratio of ~75:25. For financial distress detection, we have used these machine-learning methods: artificial neural network, categorical boosting (CatBoost), convolutional neural network, decision tree, discriminant analysis, extreme gradient boosting machine (XGBoost), extreme learning machine, logistic regression, naive bayes, random forest (RF), support vector machine (SVM). Different metrics are used for the evaluation of the methods (accuracy, AUC, specificity, sensitivity, F-score, etc).

Acknowledgements: We wish to thank Viktoras Vaitkevičius from Baltfakta for providing data and fruitful discussions.

Visualization of Zetabulbs: Mandelbulbs Associated With the Riemann Zeta Function

Lukas Kuzma, Martynas Sabaliauskas, Igoris Belovas

Institute of Data Science and Digital Technologies
Vilnius University

martynas.sabaliauskas@mif.vu.lt

Amidst the rich tapestry of the natural world, we encounter a myriad of imperfect geometric shapes. Geometry equips us with potent tools to articulate and formalize the innate symmetries governing these shapes. These fundamental symmetries encompass rotation, translation, and reflection, providing us with a framework to comprehend the harmony in nature, from intricate snowflake-like patterns to bilateral symmetry seen in many creatures.

We uncover an intriguing subset of natural phenomena defying traditional categorization within these established symmetries. For instance, consider the intricate branching of trees, which lacks rotation, translation, or reflection symmetry. Yet, zooming in on a single branch reveals a striking resemblance to the entire tree, a phenomenon known as scale symmetry. This concept extends across various natural structures, from Romanesco cauliflower spirals to mountain silhouettes, river meanders, and coastal contours.

In mathematics, we honor these scale-symmetric structures called fractals. While some fractals, like Sierpinski's triangle or Menger's sponge, ignite our imagination and lend themselves to artistic expression, others pose computational challenges. A prominent example, the Mandelbrot set exhibits an astonishingly intricate boundary, captivating mathematicians for decades, as evidenced by numerous scientific publications in the Clarivate Analytics Web of Science database.

The Mandelbrot set's significance transcends mathematics; it finds practical application in computer graphics, creating realistic and fantastical landscapes for visual storytelling and cinema.

Further exploration of fractals introduces the Mandelbulb, a 3D counterpart to the Mandelbrot set. Due to the non-existence of a 3D analog of the 2D space of complex numbers, the absence of a canonical three-dimensional Mandelbrot set led to new techniques employing spherical coordinates to construct the Mandelbulb. Moreover, by combining fractals, figures, and surfaces, we may produce novel hybrid fractal structures.

Efficiency of YOLOv5 Models in the Detection of Construction Details

Tautvydas Kvietkauskas, Ernest Pavlov, Pavel Stefanovič

Vilnius Gediminas Technical University

tautvydas.kvietkauskas@stud.vilniustech.lt

Object detection today is widely used in different areas, for example, medicine, industry, business, and even everyday solutions. New models of object detection are constantly developed and the old models are improved by adding some new features or changing the architecture of models. One of the most used object detection models in scientific research is YOLO. In recent years, some new versions of YOLO have been proposed, but they are not fully investigated and lack scientific research. The most stable version of the YOLO group algorithm is the YOLOv5. In this research, the newly collected construction detail dataset has been prepared. Images from different angles of construction details have been taken, and the dataset has been labelled. The dataset consists of 22 construction details. The experimental investigation has been performed using five different models of YOLOv5: n, s, m, l, x. During the experimental investigation, various parameters have been used to find out the influence of the parameters on the final detection results. The models have been tested on three different backgrounds: white, neutral, and mixed. The results of the experimental investigation are promising, and in the future, the models can be used in construction recommendation models.

Population-Based Algorithm for Discrete Facility Location With Ranking of Candidate Locations

Algirdas Lančinskas[1], Julius Žilinskas[1],
Pascual Fernández[2], Blas Pelegrín[2]

[1] Vilnius University
[2] University of Murcia, Spain

algirdas.lancinskas@mif.vu.lt

Facility location problems are mathematical optimization problems that involve finding the best locations for facilities (e.g., factories, warehouses, stores) to serve customers within a given geographic area. The goal is typically to minimize costs, maximize efficiency, or optimize other objectives. Facility location problems can vary in several ways, including customer behavior rules, the type of location space, which can be continuous or discrete, constraints on locations for facilities, and much more. These variations impact the complexity of the problem and the appropriate solution methods.

This research is focused on the discrete competitive facility location problem for an entering firm, which is important to firms that are entering a market and need to choose optimal locations for their facilities from a predefined set of candidate locations. Importantly, the firm must consider. The goal is to maximize the market share obtained by the new locations, considering the competition from other facilities owned by other firms in the market.

A new heuristic algorithm based on the ranking of location candidates and handling of the population of the best solutions found so far has been developed and applied to solve this facility location problem. The algorithm extends its precursor, based on a single-agent random search with the ranking of candidate locations, by including strategies to handle the population of the best solutions found so far and new strategies for ranking candidate locations, considering features of the solutions in the population.

The developed algorithm has been experimentally investigated by solving the discrete competitive facility location problem with the Pareto-Huff customers behavior rule, considering that the whole buying power of customers is proportionally divided among facilities that are Pareto optimal by distance and attractiveness.

A Hybrid of Bayesian-Based Global Search With Hooke-Jeeves Local Refinement for Multi-Objective Optimization Problems

Linas Litvinas

Vilnius University

linas.litvinas@mif.vu.lt

The proposed multi-objective optimisation algorithm hybridises random global search with a local refinement algorithm. The global search algorithm mimics the Bayesian multi-objective optimisation algorithm. The site of the current computation of the objective functions by the proposed algorithm is selected by randomised simulation of the bi-objective selection by the Bayesian-based algorithm. The advantage of the new algorithm is that it avoids the inner complexity of Bayesian algorithms. A version of the Hooke–Jeeves algorithm is adapted for the local refinement of the approximation of the Pareto front. The developed hybrid algorithm is tested under conditions previously applied to test other Bayesian algorithms so that performance can be compared. Other experiments were performed to assess the efficiency of the proposed algorithm under conditions where the previous versions of Bayesian algorithms were not appropriate because of the number of objectives and/or dimensionality of the decision space.

Cryptocurrency Price Prediction Model Development Using Machine Learning Algorithms

Gita Maliukaitė, Mantas Vaitonis

Vilnius University

gita.maliukaite@knf.stud.vu.lt

Cryptocurrencies are digital currencies that exist in a virtual space and every year more and more financial institutions do include them in their portfolio. These digital currencies are not controlled by central banks. Investors are increasingly interested in them because the their highly volatile prices, which lead to a higher return potential. These are the main factors why cryptocurrencies are driving the development of a price prediction model. Machine learning is one of the tools that allows the development of this model for cryptocurrencies. The outcome of this study is to apply machine learning algorithms to develop an accurate price prediction model for BTC and ETH currencies. To achieve this goal, several machine learning algorithms as Long Short – Term Memory and Gaussian process regression are used. Comparing these algorithms allows us to find the best price prediction model for the selected cryptocurrencies. In order to assess the accuracy of the model, the evaluation is based on the root mean square error, the mean percentage error and the R2 criterion. This study presents the methodology, data preparation and analysis, model training and results

.

Automatic Simplification of Lithuanian Administrative Texts: Initial Experiments

Justina Mandravickaitė, Danguolė Kalinauskaitė, Danguolė Kotryna Kapkan, Eglė Rimkienė, Tomas Krilavičius

Vytautas Magnus University

justina.mandravickaite@vdu.lt

We present the first steps towards the automatic simplification of Lithuanian administrative texts by transforming their language into plain Lithuanian language. Plain language is a variant of standard language that aims to present information as clearly and efficiently as possible, avoiding complex structures and constructions, as well as rare vocabulary and professional jargon in order to make texts easy to read and understand for non- specialists or experts (Maaß 2020; Nord 2018). It is mostly used for written communication of public institutions with the public, and in some countries, this practice is even required by law. Our presentation is twofold: first, we introduce the Lithuanian corpus for text simplification tasks, and second, we share the results of the initial experiments in text simplification for the Lithuanian language using the poor man's approach, which is popular in low-resources scenarios, in order to establish a strong baseline. The corpus consists of more than 2000 sentence (original and simplified) pairs. Original sentences were collected from a variety of websites of public institutions, providing social services (e.g., hospital, police). In our work, we present the text simplification guidelines we used for the corpus preparation, the process of manual text simplification, as well as the first results of qualitative and quantitative analysis of the prepared corpus. The qualitative analysis involved examining consistency, systematicity, and logic in applying text simplification rules, while the quantitative analysis included the calculation of several metrics (BLEU, ROUGE, SARI, and ROLD (Replace-only Levenshtein distance)), as well as statistical analysis of text simplification operations (words inserted, deleted, and reordered). The poor man's approach that we used for the initial experiments in text simplification covered translat-

ing data from a low-resource language (Lithuanian) into a high-resource language (English) using machine translation tools. Subsequently, we applied pre-existing models, originally designed for the English language, to analyze the translated data. Three existing models for text simplification tasks were evaluated: Scientific Abstract Simplification, SAS (Wang, 2023), Keep it Simple, KiS (Laban et al., 2021), and CTRL44 (Cripwell et al., 2022). For evaluation, four metrics were used: BLEU, ROUGE, SARI, and ROLD. CTRL44 model demonstrated better performance with a BLEU score of 0.3383, ROUGE-1 score of 0.6987, ROUGE-2 score of 0.5127, and ROUGE-L score of 0.6642, compared to the other two models. Moreover, CTRL44 had a lower ROLD Average of 121.7001, suggesting fewer changes from the original text, and a higher SARI score of 39.0852, indicating better simplification efficiency. Our future plans include developing a methodology for the automatic simplification of Lithuanian administrative texts by using the developed corpus and results of establishing a strong baseline via the poor man's approach.

Cluster-Based Classification of Consensus Protocols for Decentralized Ledger Technology

Marco Marcozzi, Ernestas Filatovas, Remigijus Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University

marco.marcozzi@mif.vu.lt

The performance and security of any Distributed Ledger Technology (DLT) solution heavily depend on the consensus protocol used. The vast variety of consensus protocols in DLT have triggered debates on how to classify them. The conventional classification of consensus algorithms relies on family (Proof of Work, Proof of Stake, etc.) or other subjective criteria. However, these classifications often place protocols with significantly distinct characteristics and performance into the same category. To overcome this challenge, this research introduces a quantitative, cluster-based classification methodology to achieve an impartial grouping of analyzed consensus protocols across various platforms. The results presented show that, using separate approaches, clustering is consistent, and it effectively separates the protocols according to their family, i.e. Proof of Work from others. An extension to this work may lead to the development of an automated tool to classify consensus protocols by means of the collected data.

The research work of E. Filatovas has received funding from the Research Council of Lithuania (LMTLT), agreement No. S-MIP-21-53.

Regional GDP Disparities in the Interwar Lithuania

Jurgita Markevičiūtė[1,2], Zenonas Norkus[1,2]

[1] Institute of Sociology and Social Work, Vilnius University
[2] Institute of Applied Mathematics Vilnius University

jurgita.markeviciute@mif.vu.lt

The research on economic cross-regional inequalities in the European Union (EU) countries serves to inform the EU cohesion policy. Recently, it was enlarged by research on the long-run trends in the changes in cross-regional inequalities. Rosés JR & Wolf N (eds) The Economic Development of Europe's Regions. A Quantitative History since 1900 (Routledge, 2019)). However, this research covers only older EU member countries. We apply the same methodology of the real GDP estimation, invented by Frank Geary and Tom Stark, to the investigation of long-run trends in the cross-regional economic productivity disparities in Lithuania. Among the three Baltic countries, Lithuania is the most challenging case for research of this kind because of the instability of its interwar borders. Poland seized by force the Vilnius region from 1920 to 1939. Lithuania was compensated in 1923 by the Klaipėda (Memel) region, ceded by Germany. In March 1939, it recaptured Klaipėda, but in October 1939, Lithuania regained Vilnius. Therefore, interwar and contemporary Lithuania are not strictly comparable as territorial units. Another challenge is that Geary and Stark's methodology best applies to census data. However, Lithuania during the interwar period had only one census in 1923. To evaluate the regional disparities during the interwar period, we apply the RAS methodology to obtain the working population by region and sector. Finally, we estimated the σ-convergence coefficients. Comparing the ratio of regional GDP per capita to the national GDP per capita, Klaipėda (Memel) was in the leading position. Still, surprisingly, Šiauliai was in second place, and the capital region, Kaunas, was only in third place. We also noted that during the last years of interwar Lithuania,

nearly at the time when the prolonged economic depression related to the global Great Interwar Recession 1929-1933 ended, divergence started to increase dramatically.

Acknowledgement: The research leading to these results has received funding from the [EEA]/ [Norway] Grants 2014-2021 for the Baltic Research Programme's project "Quantitative Data about Societal and Economic Transformations in the Regions of the Three Baltic States during the Last Hundred Years for the Analysis of Historical Transformations and the Overcoming of Future Challenges (BALTIC100)".

Bridging the Gap: Estimation of Soluble Compounds and Protein Concentrations in E. Coli Bioprocesses

Deividas Masaitis, Renaldas Urniežius, Arnas Survyla

Department of Automation
Kaunas University of Technology

renaldas.urniezius@ktu.lt

In the quest to improve the bioprocess efficiency of E. coli, an accurate estimation of the concentrations of soluble compounds is crucial. Our study presents a hybrid model approach, which takes advantage of off-gas analysis data and physiological parameters, including biomass age and specific growth rate, to estimate soluble compounds such as acetate and glutamate in fed-batch cultures. This methodology has proven to be effective in this context. Based on the success of this approach for estimating soluble compounds, we extend its application to the realm of protein concentration estimation within E. coli bioprocesses. Using a hybrid recurrent neural network, we establish intricate relationships between these parameters. Our model incorporates ensemble averaging, information gain and diverse model inputs, thus creating a more robust representation of E. coli bioprocess dynamics.

Translation of ISO/IEC 22989:2022 "Artificial Intelligence Concepts and Terminology" Standard Into Lithuanian

Saulius Maskeliūnas

Institute of Data Science and Digital Technologies
Vilnius University

saulius.maskeliunas@mif.vu.lt

Artificial Intelligence (AI) has become a topical field of Science and Technology, evolving fast globally and in Lithuania. Along with developing AI systems, international standards for Artificial Intelligence are also created. Since 2017, AI standards have been, and continue to be, prepared by ISO/IEC JTC 1/SC 42 "Artificial Intelligence" – the 42nd Subcommittee of the 1st Joint Technical Committee of the International Organization for Standardization and the International Electrotechnical Commission [1]. Lithuanian researchers/ institutions participate in ISO/IEC JTC 1/SC 42 activities through the Technical Committee TK4 "Information Technology" of the Lithuanian Standards Board [2]. In addition, members of the AI Section of the Lithuanian Computer Society [3] can learn about the draft ISO/IEC JTC 1/SC 42 Artificial Intelligence standards under consideration because the Lithuanian Computer Society is a TK4 member.

As Wael William Diab, SC 42 Chair, has announced at the SC 42 Plenary on 16th October 2023, ISO/IEC 22989:2022 "Information technology – Artificial intelligence – Artificial intelligence concepts and terminology," the first of AI standards, is now freely available [4]. It defines the concepts and terms most important to everybody interested in Artificial Intelligence.

Supporting the full use of the Lithuanian language in the most important new fields of Science and Technology, accepting by consensus a common Lithuanian terminology is of great significance. Lithuanian AI terminology adopted by consensus is necessary for Lithuania's academic community, business enterprises, and even schools: pupils have already started to learn about Artificial Intelligence as part of the updated General Curriculum Framework since September 2023.

After consideration, the working group of researchers from the Kaunas University of Technology, Vytautas Magnus University, VILNIUS TECH, and Vilnius University have decided to prepare the Lithuanian translation of this standard. The draft version of the Lithuanian translation of the ISO/IEC 22989:2022 Artificial Intelligence main terms will be laid out for discussion in the presentation.

[1] Standards by ISO/IEC JTC 1/SC 42 “Artificial intelligence” https://www.iso.org/committee/6794475/x/catalogue/

[2] Technical Committee TK 4 “Information technology” of the Lithuanian Standards Board https://eshop.lsd.lt/public#!/committee/info/6040

[3] Artificial Intelligence section of the Lithuanian Computer Society https://www.liks.lt/liks-sekcijos/liks-intelektikos-sekcija/

[4] ISO Publicly Available Standards https://standards.iso.org/ittf/PubliclyAvailableStandards/

Automatic Tumor Identification Using Deep Neural Network

Edita Mažonienė, Mantas Kundrotas, Dmitrij Šešok

Vilnius Gediminas Technical University

edita.mazoniene@vilniustech.lt

The most recent advancements in artificial intelligence (AI), particularly machine learning (ML), have made it possible to create automated solutions that, when it comes to interpreting health data, can either eliminate or significantly minimize human error. Pathologists examine slides of histopathologic tissues under a microscope, due to ethical concerns about the use of AI in pathology and laboratory medicine. This process is required by the law and cannot be replaced; pathologists are solely responsible for the outcome. Nonetheless, many automated systems could handle difficult issues requiring a very quick response time and precision, or they could work on activities requiring a quick response time and accuracy at the same time. These ML based systems can be customized to operate with medical imaging data, which enable physicians to evaluate a greater number of patient cases in less time and provide the capacity to detect early signs of cancer or other diseases, thereby enhancing health monitoring approaches. Our study's major objective was to increase the capacity to identify more precise ML approaches and procedures that may be used to identify tumor-damaged tissues in histopathological whole slide images. The trials we ran showed that the training and test datasets differed by 1% in terms of AUC. During multiple training cycles, the U-Net model achieved nearly double the model size reduction and improved accuracy from 0.95491 to 0.95515 AUC. Properly trained convolutional models performed well on different sizes of groups. The result improved to 0.96870 with the test time augmentation (TTA) method and to 0.96977 when the multi-model ensemble was added. We also discovered that by employing specific analysis approaches, it is possible to identify and correct errors in the models. The image processing parameters needed to be adjusted to increase the AUC by approximately 0.3%. Finally, after more training data preparation, the individual model's performance improved to 0.96664 AUC, which was more than 1% better than the previous best model.

Association of Genetic Variants With Myocardial Mechanics and Morphometry in Patients With Nonischemic Dilated Cardiomyopathy

Karolina Mėlinytė-Ankudavičė[1], Marius Šukys[1], Gabrielė Kasputytė[2,3], Ričardas Krikštolaitis[2,3], Paulius Savickas[2,3], Eglė Ereminienė[1], Grytė Galnaitienė[1], Vaida Mizarienė[1], Gintarė Šakalytė[1], Tomas Krilavičius[2,3], Renaldas Jurkevičius[1]

[1] Lithuanian University of Health Sciences
[2] Vytautas Magnus University
[3] Centre for Applied Research and Development

gabriele.kasputyte@vdu.lt

In the management and monitoring of patients, it is crucial to carefully analyze the relationship between phenotypic alterations in the heart and gene variations found in dilated cardiomyopathy (DCM). Dilatation and dysfunction of the left or biventricular ventricles, without aberrant loading conditions or coronary artery disease, are the hallmarks of DCM. Our study aims to evaluate the connection between cardiac-related gene variants and myocardial mechanics and morphometrics in nonischemic dilated cardiomyopathy (NIDCM) patients. This study includes NIDCM patients who underwent genetic testing via Illumina NextSeq 550 and a gene capture panel of 233 genes. We compare clinical, echocardiographic, and MRI parameters between patients with and without gene variants. Among 95 NIDCM patients, GATAD1, LOX, RASA1, KRAS, and KRIT1 genes exhibit notable differences. KRAS and KRIT1 correlate with poorer heart mechanics and chamber enlargement, while GATAD1, LOX, and RASA1 related to improved cardiac function and morphology. We conclude that these novel gene variants could impact clinical and cardiac parameters in NIDCM.

When Exponential Growth Bias Gets Worse: An Experiment With an Educational Intervention

Gerda Ana Melnik-Leroy, Linas Aidokas, Gintautas Dzemyda, Giedrė Dzemydaitė, Virginijus Marcinkevičius, Danguolė Melnikienė, Vytautas Tiešis, Ana Usovaitė

Institute of Data Science and Digital Technologies
Vilnius University

gerda.melnik@mif.vu.lt

Exponential growth is a pervasive phenomenon found in a wide array of fields, from biology, where it underpins the proliferation of microorganisms, to economics, where it relates to compounding interests, and even physics, where it plays a role in phenomena like nuclear chain reactions. Despite this, there's a growing body of literature that highlights the challenge people face in accurately grasping this type of growth. Specifically, individuals tend to consistently underestimate exponential growth and often perceive it as linear. Recent studies have attempted to explore the origins of this bias and mitigate it by employing logarithmic and linear scales in graphical representations. However, these investigations have yielded conflicting results regarding which scale induces more perceptual errors. In the present study, we conducted an experiment involving a brief educational intervention to further investigate the factors that modulate the exponential bias in graphical representations. Our hypothesis suggests that each scale may induce misperceptions in particular contexts. We also delved into the impact of mathematical education by testing two groups of participants, one with a humanities background and the other with a background in formal sciences. A pretest-posttest design was chosen for the study.

The results of our study confirmed that when employed in an inappropriate context, both logarithmic and linear scales can negatively affect the interpretation of visualizations representing exponential growth. Specifically, the log scale leads to more errors in tasks related to describ-

ing graphs, while the linear scale misguides individuals when making predictions about the future trajectory of exponential growth. The latter part of our study reveals that these challenges with both scales can be mitigated through a brief educational intervention. Importantly, while no difference was observed between the two participant groups before the intervention, those with a stronger mathematical education displayed more significant learning effects during the posttest. We discuss the findings of this study in the theoretical framework of the dual-process model.

Machine Learning Approaches for the Analysis of Functional Brain Connectivity Patterns in Depression

Gajane Mikalkėnienė[1,2], Jolita Bernatavičienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Republican Vilnius Psychiatric Hospital

gajane.mikalkeniene@mif.stud.vu.lt

According to recent statistics, in Lithuania, in 2022, there were 23,97 per 1,000 residents who suffered any kind of depression. Depression is one of the suicide risk factors, and Lithuania is known for its high suicide rate (first place in the European Union, seventh worldwide). Improvement of prevention and treatment of depression using artificial neural networks could help to lessen illness severity and improve treatment efficacy, which is crucial for improving life quality, mental health, and well-being.

During mental disorders, there are changes in behaviour, but also in brain bioelectric activity, which could be monitored using neuroimaging methods such as electroencephalography, magnetoencephalography, and functional magnetic resonance imaging. Data acquired using those methods can be analyzed as functional brain connectivity. To understand this concept better and to find out if connectivity can be a sustainable biomarker of depression or other mental disorders, several machine learning methods were applied to classify patients and healthy subjects, to distinguish depressed patients from patients with other mental disorders, to predict the symptom change and of treatment efficacy.

The purpose of this research was to evaluate machine learning methods used in this field, what results were achieved, and what are strengths and limitations by analysis of related articles found using the PubMed database from 2000 to 2023. This analysis is important for potential enhancements to these methods, leading to improved diagnostic and therapeutic approaches for depression.

A Causality Space Model of the Web Service Quality Based on Fuzzy Theory

Jolanta Miliauskaitė[1], Diana Kalibatienė[2], Asta Slotkienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Vilnius Gediminas Technical University

diana.kalibatiene@vilniustech.lt

The quality of Web Services (QoS) is an essential characteristic in selecting a web service (WS) and achieving appropriate results regarding end-user expectations and satisfaction. In the scientific literature, authors have proposed various QoS attributes, like throughput, latency, response time, etc., that allow us to determine the WS QoS at different software systems development layers, such as business service layer, business process layer, WS layer, component layer, infrastructure service layer, and network layer. All these layers directly and indirectly influence each other. Therefore, we need an approach describing and allowing us to determine the causality relationships among QoS attributes in different layers. Understanding and modelling those causality relationships allows us to improve the WS QoS, its internal validity, and the robustness of WS. In this research, we present the causality space model that identifies QoS attribute relationships at different layers and models them using a Fuzzy Set Theory. The proposed WS QoS causality space model allows the researchers and practitioners to view and deeper understand the WS development peculiarities and the end-users to select the most suitable WS.

Development of a Modern Forest Decision Support System for Lithuania: Simulation or Optimization?

Gintautas Mozgeris[1], Arnas Matusevičius[1,2], Gabrielė Kasputytė[1,2], Ljusk Ola Eriksson[3], Tomas Krilavičius[1,2], Laimonas Butkus[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development
[3] Linnaeus University, Sweden

arnas.matusevicius@vdu.lt

Forest ecosystems and forestry in Lithuania are encountering numerous challenges, including the increased use of timber and bioenergy, uncertainties stemming from climate change, the dynamics of evolving global markets, rural development, and growing public interest in the social values of forests. Effectively managing such a multifaceted system is only possible through the implementation of smart technologies, with a growing emphasis on forest decision support systems. Historically, forestry decision support in Lithuania has relied on forest growth and forestry simulators, such as Kupolis. The application of forest simulators typically involves modelling the future development of forest resources under various conditions and specifications that describe the behaviour of the forest and forest managers.

This research aims to introduce a prototype of a modern forest decision support system that focuses on optimizing forest management alternatives to achieve specified long-term management objectives and constraints. This new software tool has been developed within the framework of the EU-financed project Forest 4.0. The initial results of optimization generally suggest shorter optimal forest rotations than the currently adopted minimal final harvesting ages, all while maintaining the sustainability principles of sustainable forest management that prioritize multiple ecosystem services. This project has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101059985.

UniDive – Universality, Diversity and Idiosyncrasy in Language Technology

Gediminas Navickas, Gražina Korvel

Institute of Data Science and Digital Technologies
Vilnius University

gediminas.navickas@mif.vu.lt

UniDive is the COST Action (CA21167) project. It is an interdisciplinary scientific network devoted to universality, diversity and idiosyncrasy in language technology.

The main objective of UniDive is to reconcile language diversity with rapid progress in language technology. It embraces both inter- and intra-language diversity, i.e. a diversity understood both in terms of the differences among the existing languages and of the variety of linguistic phenomena exhibited within a language. It gathers about 250 interdisciplinary experts (linguists, computational linguists, computer scientists, psycholinguists, and industrials) from almost 40 COST countries. It represents dozens of languages from many different language generation perspectives.

Project structure consists of four working groups: 1) Corpus annotation, 2) Lexicon-corpus interface, 3) Multilingual and cross-lingual technology and 4) Quantifying and promoting diversity.

The project is open for new members and participants. Become UniDive member, participate in Training schools, Short Term Scientific Missions (STSM), Joint publications, Grants for young researchers.

Railway Track Vibration Analysis and Complexity Assessment Based on H-ranks

Ugnė Orinaitė, Minvydas Ragulskis

Kaunas University of Technology

ugne.orinaite@ktu.lt

This study addresses safety concerns in rail transport, with a specific focus on rail vehicle movement. Railway level crossings pose the highest accident risk due to the interaction of road and rail traffic. The most crucial safety aspect at unprotected railway crossings is the driver's ability to see an approaching train. This study aims to mitigate these issues by developing a method to provide advance train warnings, regardless of visibility limitations.

Signal analysis plays a pivotal role in various scientific and technical domains, such as transportation engineering. The H-rank method is introduced as a potential tool for signal analysis. The H-rank algorithm leverages advanced mathematical concepts, particularly matrix factorization and rank estimation, to extract valuable insights from signals. It enables the identification of underlying patterns, anomalies, and pertinent aspects within the signal. As an algebraic mathematical feature, the H-rank method can not only determine the order of linear recurrence but also assess the algebraic complexity of time series. This study examines three different types of experimental train track vibration signals generated by various train types; all signals have a common sampling rate. The study uses H-ranks as the method for vibration signal analysis, the reconstructed linear regression model takes place to indicate approaching trains.

The method presented in this study relies on rail vibration measurements and can make independent predictions, not relying on other rail transport information systems. This study aims to showcase the extensive applicability of the H-rank approach in signal processing. It illustrates its capacity to reduce signal noise, enhance signal quality, and detect real-time signal variations. The outcomes of employing the H-rank algorithm underscore its advantages over conventional signal analysis techniques.

New Therapeutic Opportunities Discovered by Statistical Models and AI Algorithms

Vytautas Ostaševičius

Kaunas University of Technology

vytautas.ostasevicius@ktu.lt

Technological advances in high-speed data processing have transformed medical biology into a field of data mining, where new data sets are regularly dissected and analysed with the help of increasingly sophisticated statistical models and artificial intelligence algorithms. One of the most common ways of identifying health status is through blood analysis. As a result of this research, faster statistical and artificial intelligence methods have been proposed for speeding up the analysis and interpretation of blood parameters, allowing to avoid the mistakes of inexperienced analysts, and to take timely actions to improve human health. Such a possibility of improving human health was revealed by affecting the blood with low-frequency ultrasound, the influence of which is associated with the exchange of O2 and CO2 gases in red blood ceells and platelet aggregation. Statistical analysis, ANOVA and the non-parametric Kruskal-Wallis method, was used to evaluate the effect of ultrasound on various blood parameters. The obtained results suggest that there are statistically significant variances in blood parameters attributed to low-frequency ultrasound exposure. Furthermore, among the five machine learning algorithms employed to predicts ultrasound's impact on platelet counts, Support Vector Regression (SVR) exhibited the highest prediction accuracy, yielding an average MAPE at 10.34%. Notably it was found that the effect of ultrasound on hemoglobin in red blood cells outperformed its impact on platelet aggregation highlighting the significance of hemoglobin in facilitating the transfer of oxygen from the lungs to the body's tissues.

Acknowledgement: This research was funded by European Regional Development Fund (project No 01.2.2-LMT-K-718-05-0076) under grant agreement with the Research Council of Lithuania (LMTLT). Funded as European Union's measure in response to Cov-19 pandemic.

Back to the Roots: On the Use and Principles of Clustering Under Various Circumstances

Jan W. Owsiński

Systems Research Institute, Polish Academy of Sciences, Poland

Jan.Owsinski@ibspan.waw.pl

The paper is divided into the following essential parts: (1) general remarks on the use and meaning of cluster analysis; (2) two examples of the use of cluster analysis, showing unexpected and highly meaningful results; (3) potential use of the so-called "ideal structures" in clustering. In the first part, the most important problems arising in the context of the use of clustering methods are indicated and discussed (formulation, distances, shapes of clusters, cluster number, computational efficiency). The second part shows the highly telling two examples of the application of clustering, one to the voting by individual MPs in the Polish Parliament (the Diet) during one of the past terms of the House, and the second – to the results of a small survey among students, carried out recently among students from different countries and schools, to which the paradigm of "reverse clustering" was applied. The third part is devoted to the analysis of one of the earliest precepts ("ideal structures"), tried out in the domain of clustering, which turned out to almost at once be generally invalid, but, at the same time, potentially leading to some effective procedures.

Keywords: clustering, clustering procedure, algorithms, reverse clustering, ideal structures.

Acquiring Knowledge for Mimicking Dysarthric Speech by Incorporating Its Features Into Synthetic Speech

Tomasz Piernicki[1], Grazina Korvel[2], Bozena Kostek[1]

[1] Gdansk University of Technology, Poland
[2] Institute of Data Science and Digital Technologies
Vilnius University

grazina.korvel@mif.vu.lt

The purpose of this study is two-fold. First of all, it is to conduct in-depth analyses, allowing the extraction of features associated with dysfunctional speech, and in particular, dysarthria. The methods of speech signal analysis, such as temporal, spectral, time-frequency, cepstral, as well as linear predictive coding (LPC), are used. In addition, linear predictive cepstral coefficients (LPCC) and perceptual linear predictive coefficients (PLP) are also investigated. Resulting from this analytical approach is a set of features that corresponds best to dysarthria. Hence, the second purpose of this study is to propose techniques that may be employed to synthesise normal speech patterns with the most relevant dysarthria features to create dysfunctional speech. Several speech synthesis techniques are applied to that end, and their outcome is examined both by objective measures and subjective tests. The results are shown in the form of stationary analyses and regarding a sequence of dysarthric utterances to highlight changes detected changes. The summary of the experiments includes conclusions and plans for future research studies related to automatic recognition of dysarthric speech.

Machine Learning-Based ChatGPT Usage Detection in Open-Ended Question Answers

Birutė Pliuskuvienė, Urtė Radvilaitė,
Pavel Stefanovič, Simona Ramanauskaitė

Department of Information Technology
Vilnius Gediminas Technical University

birute.pliuskuviene@vilniustech.lt

Today, chatGPT is one of the most widely used large language models for different tasks. The open access to the possibilities of the chatGPT leads to problems in the educational area because many students start to use it to solve various knowledge assessment tasks: homework, tests, midterms, and exams. In such cases, students start cheating instead of trying to understand the study material. In this research, the suitability of traditional machine learning algorithms to detect the usage of chatGPT in the answers text of the open-ended questions was examined. The dataset was collected using the midterm exam answers of the VILNIUS TECH bachelor's students. The experimental investigation has been performed by dividing the dataset into three (student answer, chatGPT answer, rephrased chatGPT answer) and two (student answer, non-student answer) classes separately. The various combinations of text pre-processing have been taken into account, and in this way, the influence of pre-processing options has been analysed for chatGPT usage detection models' accuracy.

Performance-Based SLO Recovery for Containerized Applications

Olesia Pozdniakova, Dalius Mažeika

Vilnius Gediminas Technical University

dalius.mazeika@vilniustech.lt

The development of cloud-ready applications involves a focus on scalability and loose coupling of the containerized microservices to guarantee seamless deployment on cloud or container orchestration platforms. Auto-scaler is a key component that is responsible for dynamic resource provisioning and QoS provided by the cloud. Premature or excessive resource provisioning may increase costs, while delays could lead to service degradation and SLA violation. Hence, finding the balance between avoiding SLA violations and effective cost management is the problem addressed by the majority of auto-scaling algorithms. Rule-based and predictive machine learning-based algorithms are used to determine the proper number of resources in order to meet SLO requirements. Usually, SLI values such as CPU utilization or transactions per second are included in datasets. However, the existing auto-scaling approaches minimize the risk of SLA violations but do not address the problem of degraded SLA i.e., they lack a mechanism for overall SLA restoring after a violation and leaves a gap in the comprehensive SLA management.

Threshold-based autoscaling approach was proposed and investigated. Auto-scaler monitors the SLO value during the particular evaluation timeframe and restores the SLO in the case of violations. The proposed autoscaling approach ensures that the application operates at an elevated performance for a specific time frame to achieve the SLO target during a specified time. An experimental study was conducted to evaluate the ability of the proposed auto-scaler to recover and maintain SLO under five distinct load scenarios. The obtained results were compared with the similar dynamic thresholds-based autoscaling solution known as DM that previously demonstrated good results in terms of resource provisioning and response time. Several criteria such as the amount of the provided resources, and the number of over-provisioned as well as

under-provisioned pods were used to assess the effectiveness of the proposed algorithm to recover and maintain the required SLO level.

The results revealed that the proposed auto-scaling solution demonstrates better adherence to SLA across the majority of evaluated load scenarios, even when employing a similar number of resources as other threshold-based algorithms.

Survival With Random Effect

Rokas Puišys, Jonas Šiaulys

Institute of Mathematics
Vilnius University

rokaspuisys@gmail.com

The presentation focuses on mortality models with a random effect applied in order to evaluate human mortality more precisely, which could be used in life insurance when compiling mortality tables. Such models are called frailty or Cox models. The main assertion shows that each positive random effect transforms the initial hazard rate (or density function) to a new absolutely continuous survival function. In particular, well-known Weibull and Gompertz hazard rates and corresponding survival functions are analysed with different random effects. These specific models are presented with detailed calculations of hazard rates and corresponding survival functions. Six specific models with a random effect are applied to the same data set of Baltic countries mortality. The results indicate that the accuracy of the model depends on the data under consideration.

Effectiveness of Machine Learning Algorithms for the Detection of Soft Tissue Calcifications in Panoramic Radiographs

Alina Pūrienė[1], Paulius Raškevičius[1], Saulė Skinkytė[1], Lina Stangvaltaite-Mouhat[3], Indrė Stankevičienė[1], Darius Padvelskis[2]

[1] Institute of Dentistry, Vilnius University
[2] MB DentFuture
[3] Oral Health Centre of Expertise in Eastern Norway, Oslo, Norway

alina.puriene@gmail.com

Introduction. Previous studies have tested the use of AI technologies for diagnosing different conditions such as caries, periodontal pathology, sinusitis, osteoarthritis, facial and jaw cysts, and tumors. However, there is a lack of data on the detection of soft tissue calcifications. **Material and methods.** The present study aimed to identify soft tissue calcifications in 1,100 panoramic radiographs of patients who visited Vilnius University Hospital Zalgiris Clinic from 2014 to 2016 for clinical purposes. The study protocol was approved by the Vilnius Regional Biomedical Research Ethics Committee. The radiographs were analyzed using Romexis Planmeca Viewer software. After annotating all the images, machine learning algorithms were developed for automatic assessment of soft tissue calcification in radiographs, using the Pynthon 3.8 programming language. Model performance was evaluated using the F1-score metric and other data-depended methods. **Results.** Detection of salivary gland calcification in the radiographs yielded an intersection over union (IoU) of 0.803 and an accuracy 90.2%. Identification of calcifications of the elongated styloid process nd carotid artery calcifications resulted in IoU of 0.7194 and an accuracy of 90%. **Conclusion.** The initial machine learning algorithms model demonstrated promising results, indicating the potential for improved diagnostic effectiveness in detecting salivary gland, elongated styloid process, and carotid artery calcifications using the appropriate application of artificial intelligence technologies. To enhance the development of the model, it is recommended to include a larger number of panoramic radiographs containing soft tissue calcifications.

Hiding Multiple Images in Coupled Lattices of Hyper Fractional Maps

Jūratė Ragulskienė

Kaunas University of Technology

jurate.ragulskiene@ktu.lt

The image hiding scheme in the two-dimensional coupled map lattice of nilpotent matrices is presented in this talk. The complexity of the hyper coupled map lattice is increased by replacing scalar nodal variables by n-dimensional iterative matrix variables. The proposed image hiding scheme is implemented using the matrix representation of the mapping function of the fractional logistic map. It is demonstarted that the spatiotemporal divergence induced by nilpotent nodal matrices can yield multiple secret images at diferent discrete moments of time during the time evolution of the lattice. The carrying capacity of the proposed scheme is n − 1 diferent dichotomous digital images, where n is the dimension of the nilpotent nodal matrices. Computational experiments are used to demonstrate the functionality of the proposed scheme.

Tree Structure-Based Competencies Formalisation: Path Towards Alignment of Industry Needs, Person Competency Portfolio and Study Program Objectives

Simona Ramanauskaitė, Pavel Stefanovič, Antanas Čenys

Department of Information Technology
Vilnius Gediminas Technical University

simona.ramanauskaite@vilniustech.lt

The study program and course orientation on gained competency leads to better alignment with industry expectations. However, the usage of natural language descriptions to define the learning outcomes and results complicates the unambitious understanding of the gained competencies. It is an important issue both for humans and mostly for computerised systems, trying to align the competencies of different students, study programs, and industry needs. To solve the issue, we adopt tree structure-based topic data, cognitive learning and disposition levels to define the competency. Developed solutions for competency formalisation are tested and indicate the data structure usage provides possibilities for automated competency comparison, study path generation, and matching between industry needs and student portfolio or study program vision.

Acknowledgement: The results were achieved in MERIT project (grant agreement no. 101083531), co-funded by the European Union.

Investigation of Speech Signal Processing Parameters in Wave-U-Net Source Separation

Justina Ramonaitė, Pooja Gore, Gražina Korvel, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University

justina.ramonaite@mif.stud.vu.lt

Separating clean speech from noise in order to enhance speech quality and intelligibility is a challenging task known as speech denoising, where the input is noisy speech. The challenge is mainly caused by non-stationary noise and low signal-to-noise ratio (SNR). Deep learning models are being increasingly used to solve this task due to their superior performance in non-stationary noisy environments compared to conventional approaches. Deep learning methods model the nonlinear relationship between clean speech and noisy speech signals without prior knowledge of noise statistics. In this study, we focus on predicting end-to-end audio source separation. We use a waveform-based method, namely Wave-U-Net, which is an adaptation of the U-Net architecture to a one-dimensional time domain. The goal of this research is to investigate the correlation between Wave-U-Net performance and speech processing parameters, including speech sampling frequency and frame length. For this purpose, the speech signal is downsampled using sampling frequencies of 8 kHz, 11.025 kHz, 16 kHz, and 48 kHz, and divided into short-time intervals ranging from 10 ms to 50 ms in 10 ms increments. The experimental results reveal the significant influence of these parameters on denoised speech quality.

Is a Decline in Reading Achievement Due to a Low Level of SES? Is It (Im)Possible?

Laura Ringienė[1], Gabrielė Stupurienė[1], Rita Dukynaitė[2], Rimantas Želvys[3], Audronė Jakaitienė[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Ministry of Education, Science and Sport
[3] Institute of Educational Sciences
Vilnius University

gabriele.stupuriene@mif.vu.lt

The Progress in International Reading Literacy Study (PIRLS) is an international assessment and research project designed to measure reading achievement at the fourth-grade level. In addition, questionnaires are given to students' teachers, school principals, and parents to gather information about students' experiences in developing reading literacy. Researchers seek to explain the reading achievements of students in PIRLS by various factors, including socioeconomic status (SES). The construction of the SES index for PIRLS is based on the approach used in PISA (Programme for International Student Assessment). Students are scored according to their parents' reports regarding the four indicators (number of books in the home; number of children's books in the home; highest level of education of either parent; highest level of occupation of either parent) on the Home Socioeconomic Status scale. The scale is divided into three categories (levels): high, medium, and low [1].

The study aims to determine whether lower reading achievement can be explained by a low level of SES. The data set consisted of 12 countries from the EU that participated in all five cycles of PIRLS (covering the period 2001-2021). For the analysis, 5 plausible values of reading achievement were used. There were no missing values for reading achievement, but this was not the case for SES. All statistical analyses were performed with each plausible value separately and averaged afterwards.

The results showed that the mean reading by SES levels was lower for the low-SES group. For example, in Lithuania, the average reading score for the low-SES group of students ranged from 503 (international avg. 443) in 2001 to 430 (international avg. 456) in 2021. This indicates a decrease in the reading skills of the low-SES group of students. Meanwhile, the average reading achievement of the high-SES group of students increased from 588 (international avg. 548) in 2001 to 604 (international avg. 542) in 2021).

However, the results show that most of the data came from the medium SES group. The proportion of students in the medium SES group ranges from 23.9% to 86.2% in the analyzed countries. Meanwhile, the proportion of low-SES groups of students varies from 0.05% to 17.2%.

Also, after analysing the data collected from the parents, it turned out that in some countries a large share of the data was omitted in each cycle. For example, omitted SES entries in England vary from 45% to 100% in each cycle, in Germany from 15.6% to 44.4%, in the Netherlands from 33.2% to 53%, and in Sweden from 7.8% to 43.9%. Only in Bulgaria did the percentage of low SES students exceed 10% at each level. The question then arises: can we interpret the data correctly if we don't have enough data for all three levels of SES?

The analysis of the PIRLS data on reading achievement by SES revealed that conclusions should be drawn very cautiously.

[1] https://pirls2021.org/results

A Health Data Analytics Maturity Model for Hospitals Information Systems

Álvaro Rocha

University of Lisbon – ISEG, Portugal

amr@iseg.ulisboa.pt

In the last five decades, maturity models have been introduced as reference frameworks for Information System (IS) management in organisations within different industries. In the healthcare domain, maturity models have also been used to address a wide variety of challenges and the high demand for Hospital IS (HIS) implementations. The increasing volume of data exceeds the ability of health organisations to process it for improving clinical and financial efficiencies and quality of care. It is believed that careful and attentive use of Data Analytics in healthcare can transform data into knowledge that can improve patient outcomes and operational efficiency. A maturity model in this conjuncture, is a way of identifying strengths and weaknesses of the HIS maturity and, thus, finding a way for improvement and evolution. This speech presents a proposal to measure Hospitals Information Systems maturity regarding Data Analytics. The outcome is a maturity model, which includes six stages of HIS growth and maturity progression.

Kernel Density Estimator by Minimizing Bias

Tomas Ruzgas, Kristina Pupalaigė

Kaunas University of Technology

tomas.ruzgas@ktu.lt

A histogram is one of the oldest and most popular density estimators. Histogram and its representation were first introduced in 1891 by Karl Pearson. For the approximation of density, the number of observations falling within the range is calculated and divided by sample size and the volume of range. The histogram is based on a step function. Derivatives, which can be equal to zero or not defined, strongly affects the further histogram analysis. For example, it can cause problems when trying to maximize a likelihood function which is defined in terms of the densities of the distributions.

It is important to mention that the histogram was kept as the only nonparametric density estimator until 1950's, while substantial and simultaneous progress was made for density and spectral density evaluations. Later in 1951, Fix and Hodges, in a not very well known publication, presented the basic algorithm of nonparametric density evaluation. This previously not published technical report was formally presented to the public only 1989, as review made by Silverman and Jones. Researchers have focused on the problem of statistical discrimination and did their investigations when the parametric form of the sampling density was not originally known. Later, several common algorithms and alternatives in theoretical modeling were introduced by Rosenblatt in 1956, Parzen in 1962, and Cencov in 1962. Then followed the second wave of important and primarily theoretical papers by Watson and Leadbetter in 1963, Loftsgaarden and Quesenberry in 1965, Schwartz in 1967, Epanechnikov in 1969, Tarter and Kronmal in 1970, and Kimeldorf and Wahba in 1971. The natural multivariate generalization was introduced by Cacoullos in 1966. Finally, in the 1970's the first papers focusing on the practical applications of these methods were published by Scott et al. in 1978 and Silverman in 1978. These and later multivariate applications awaited the computing revolution.

Since the kernel estimate is calculated at sample points, a bias occurs, the effect of which is strongly felt in small sample sizes. Various techniques are used to reduce its influence. Here, our approach was slightly different. We construct a kernel function whose form is such that the influence of observations on the estimation is reduced, and the main attention is placed on their environment. The form of the proposed kernel function is complex, which in turn raises other challenges.

Contemporary Approaches to Investment Portfolio Decision-Making

Darius Sabaliauskas, Jolanta Miliauskaitė

Vilnius Gediminas Technical University

dariussabaliauskas@stud.vilniustech.lt

H. M. Markowitz's groundbreaking Modern Portfolio Theory (MPT), introduced in 1952, started a new era in the investment domain. MPT's main concepts and principles are Diversification, Efficient Frontier, Risk and Return, Expected Return and Variance, Correlation, Portfolio Optimisation, Risk-Free Asset, Capital Market Line, and Security Market Line. MPT has significantly influenced the finance industry, reshaping how investors, portfolio managers, and financial analysts approach asset allocation and risk management. With the ongoing emergence of the fourth industrial revolution, characterized by significant advancements in computing and artificial intelligence, arose new types of machine learning, neural networks like Recurrent Neural Networks, more precisely, the Long Short-Term Memory (LSTM), the Gated Recurrent Unit (GRU) or fuzzy logic, the investment sector has witnessed others continuous changes. New techniques and methods like meta-heuristic algorithms (Tabu search, swarm approaches, evolutionary algorithms) and mathematical algorithms constantly emerge each year to format investment portfolios to maximize profitability while minimizing risk. This study unveils and analyses the most recent methodologies in crafting investment portfolios over the past five years [2017-1th November of 2023]. The study findings show these contemporary approaches and contribute to improving the existing knowledge base surrounding constructing investment portfolios. In this area, there is a lack of research that combines financial engineering with computer science. Therefore, the study could be of interest not only to researchers but also to other stakeholders, such as financial brokers, investment funds, and private investors, to view and better understand the current state of existing methods and models for analyzing and forming the investment portfolio.

Customer Segmentation for Personalised E-Commerce Advertising Campaigns

Virgilijus Sakalauskas, Dalia Krikščiūnienė

Vilnius University

virgilijus.sakalauskas@knf.vu.lt

The continuous rise of e-commerce has encouraged significant interest among researchers in comprehending online shopping behavior, consumer interest trends, and the effectiveness of advertising strategies. However, a notable research gap exists in the identification of promising e-shoppers for tailored advertising campaigns. In response, this paper introduces an innovative approach to identify high-value e-shop clients through the strategic analysis of clickstream data. Our novel algorithm is designed to determine customer engagement and make out high-value customers. It uses clickstream data to compute a Customer Merit (CM) index, which evaluates the customer's engagement level and anticipates their purchase intent. The CM index dynamically adapts, taking into account the customer's activity, efficiency in product selection, and time spent browsing. This approach proves its value to businesses aiming to identify potential buyers and optimize e-shop sales through cost-effective advertising campaigns. To validate our approach, we tested it with actual clickstream data from two e-commerce websites. The results demonstrate that our personalized advertising campaign outperformed the non-personalized counterpart, as evidenced by improved click-through rates and conversion rates. Emphasis is placed on the method's efficacy in identifying potential e-shop visitors with a high purchase intent. This methodology integrates customer browsing and purchasing behaviors with key metrics such as time spent on the website and frequency of e-shop visits. In summary, our findings underscore the potential of personalized advertising strategies in augmenting e-commerce sales while simultaneously reducing advertising costs. By harnessing the power of clickstream data and adopting a targeted approach, e-commerce businesses can not only attract but also retain high-value customers, leading to increased revenue and profitability.

Reinforcement Learning Based Model for Personalising the Picture Exchange Communication System

Asta Slotkienė, Augustas Mikulėnas

Vilnius Gediminas Technical University

asta.slotkiene@vilniustech.lt

The ability to communicate effectively with children diagnosed with autism spectrum disorders (ASD) is an essential skill for adults. To facilitate the communication, researchers have proposed augmentative and alternative communication systems, such as the Picture Exchange Communication System (PECS). To improve their usability, the systems are digitalised. Artificial Intelligence (AI) capabilities are applied to increase communication effectiveness. Most existing research works based on AI models try to build classifiers for the child's reaction states, assuming that the data to train the models are fully labelled. However, data labelling is prone to subjective interpretations by the specialist. For each child with ASD, their individuality makes it impossible to determine the rule-based data and their decisions. In this paper, we propose a strategy learning model, where the agent uses reinforcement learning (RL) to learn an optimal strategy of presenting an appropriate PECS card for effective communication between a child with ASD and an adult. The study included experiments with different RL environments and their parameters, describing different sets of states and prescribing additional metadata to find the optimal learning strategy. The main result is the proposed new RL-based model, which allows the achievement of the goal with the minimal number of steps by reducing the count of episodes and utilising a strategy that maximises the probability of achieving the goal.

Using Deep Learning and Visual Analytics to Investigate Hate Speech Patterns in Lithuanian Politics

Milita Songailaitė[1,2] , Justina Mandravickaitė[1,2],
Justinas Juozas Dainauskas[1]

[1] Department of Informatics, Vytautas Magnus University
[2] Centre for Applied Research and Development

justinas.dainauskas@vdu.lt

Hate speech, a multifaceted and significant phenomenon, plays a pivotal role in inciting violence rooted in factors like race, ethnicity, sexual orientation, social standing, and religion. Such speech specifically targets individuals or groups based on distinguishing attributes, encompassing gender, nationality, language, and religious beliefs, among others. By molding and propagating opinions anchored in stereotypes, hate speech not only fosters violence against susceptible societal factions but also seeks to diminish and degrade them. The Lithuanian Parliament (LR Seimas) and its members stand as potent influencers, shaping both national policies and public perceptions on a myriad of subjects. This research delves into the prevalence of hate speech within the transcripts of the Lithuanian Parliament sessions. Our comprehensive analysis spans various perspectives, including individual parliamentarians, party factions, the governing body, and the opposition. Notably, we spotlight topics that elicited the most pronounced instances of hate speech. Our investigative approach combines artificial intelligence techniques, specifically a model fine-tuned for Lithuanian hate speech detection, with visual analytics, translating raw data into insightful visual representations. Our findings revealed that 2.87% of the speeches were hate-driven and 2.05% were offensive in nature. Interestingly, opposition factions exhibited a higher frequency of hate and offensive speeches compared to government factions. Though our AI-driven methodology provides insights into hate speech patterns, it's essential to understand that its classifications are based on predefined parameters, and we present these findings without political bias, emphasizing that our intent is not to cast aspersions but to contribute to academic discourse.

Deep Learning Approaches to Detect Disinformation Across News Platforms and Social Media

Milita Songailaitė, Justina Mandravickaitė,
Eglė Rimkienė, Anton Volčok, Tomas Krilavičius

Vytautas Magnus University
Center for Applied Research and Development

milita.songailaite@card-ai.eu

The digital landscape is rife with fake news, challenging democratic values, reputations, and societal trust. Incidents, like COVID-19 misinformation and drug-related hoaxes in the US and Brazil, accentuate the need for reliable detection mechanisms. We investigated deep learning techniques for disinformation detection, focusing on the Latent Dirichlet Allocation (LDA), Bidirectional Encoder Representations from Transformers (BERT), and Robustly Optimized BERT Pretraining Approach (RoBERTa). Using a comprehensive dataset derived from news articles sourced from online platforms such as The Guardian, New York Times, Sputnik, TASS, as well as various social media messages, our topic modeling revealed distinct topics including the Russo-Ukrainian war, sanctions on Russia, nuclear and chemical warfare, and terrorism. Notably, a significant portion of disinformation was identified within the Russo-Ukrainian war topic. Initial evaluations of the disinformation detection showed DistilBERT achieving an accuracy of 0.79 and RoBERTa at 0.74. After fine-tuning, RoBERTa's performance surged, reaching an impressive accuracy of 0.99, outperforming DistilBERT's 0.96. RoBERTa's precision, recall, and F1-score also surpassed DistilBERT, solidifying its superiority in our tests. It's essential to note that our collected data sample was relatively small. In future work, we plan to collect a larger sample of potential disinformation data to further validate and enhance our findings.

Would It Be Possible to Optimise the Secondary ICT Term Creation in Latvian?

Dace Šostaka, Juris Borzovs, Jānis Zuters

Faculty of Computing
University of Latvia

dace.sostaka@lu.lv

The Information and Communication Technologies Sub-commission of the Terminology Commission (ICTSTC) of the Academy of Science of Latvia was founded in 1992; we work during the academic year, meeting every fortnight.

In general, the process of secondary term creation (STC) has three parts:

1. Receiving information and communication terminology (ICT) terms in English from various sources (the European Commission, State Language Centre, ISO standards and others).
2. Discussing the terms within the framework of ICTSTC to create corresponding Latvian ICT terms.
3. Disseminating the accepted terms.

Our presentation will focus on the second part, namely, on the process of discussion to create Latvian ICT terms.

In particular, the case study of the ICT primary term "dependability" will be analysed, searching for the Latvian secondary ICT term. It will reflect the step-by-step process of considering various pros and cons when accepting each possible analogue as a Latvian secondary term.

Two concluding remarks regarding further research of the theme. Would it be possible to apply artificial intelligence to the task of optimising the terminology process in general? How exactly, most constructively, could artificial intelligence be used to optimise the Latvian term creation process in particular?

Detection of Pancreatic Cancer on CT Images Using Pseudo-Labelling Methods

Aušra Šubonienė[1], Olga Kurasova[1], Gintautas Dzemyda[1], Viktor Medvedev[1], Aistė Gulla[2], Artūras Samuilis[3], Džiugas Jagminas[3], Kęstutis Strupas[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Institute of Clinical Medicine, Faculty of Medicine
Vilnius University
[3] Institute of Biomedical Sciences, Faculty of Medicine
Vilnius University

ausra.suboniene@mif.vu.lt

This study addresses the issue of pancreatic cancer detection through the classification of computed tomography (CT) images using a semi-supervised deep learning system. Annotating medical computed tomography images is a resource-intensive process, as it requires medical experts to label each image individually within a CT scan sequence. Given the scarcity and challenges associated with acquiring labelled data, coupled with the high demand for extensive labelled datasets by contemporary machine learning techniques, the generation of pseudo-labels for unlabeled data serves to augment the volume of accessible data for machine learning applications, thereby potentially enhancing their overall performance. The aim of this research is to explore various pseudo-labelling methods to expand the labelled dataset. The study employs probability-based, entropy-based, and proximity-based pseudo-labelling methods to generate additional pseudo-labels within the unlabelled dataset. To assess the impact of pseudo-labelling on classification accuracy, both public and Vilnius University Hospital Santaros klinikos CT image datasets are utilised in this research. The convolutional neural network was employed to assess the influence of pseudo-labels on pancreatic cancer detection outcomes. The utilisation of diverse data not only enhances the reliability of the results but also allows for an evaluation of the quality of pseudo-labelling in different patient groups. By combining publicly available and private datasets, this study seeks to provide a comprehensive evaluation of the effectiveness criteria for pseudo-labelling.

Automated Hypertension Detection Using Convolutional Neural Networks

Kristupas Takšelis, Liepa Bikulčienė, Eglė Butkevičiūtė

Kaunas Faculty
Vilnius University

liepa.bikulciene@knf.vu.lt

Cardiovascular diseases, notably hypertension, pose a significant threat to global public health, contributing substantially to mortality rates. Timely diagnosis and intervention are crucial in preventing the adverse consequences of heart disorders, including damage to vital organs. The primary objective of this study is to develop a robust ECG-based methodology for heart disease detection and classification, with a specific focus on distinguishing between hypertensive and healthy patients. The ECG signals were carefully prepared by removing noise, and correcting baseline wander. Furthermore, the heartbeats were automatically identified and isolated from the ECG signal data. These segmented heartbeats were plotted on a 480x480px image and classified using Convolutional Neural Network (CNN). In this study, the open source SHAREE and PTB databases were utilized, which include 139 hypertensive and 52 healthy patients respectively. Created model was evaluated on real-world data taken from hypertensive and healthy patients. Our model successfully classified patients with higher than 96% accuracy. This study proposes a novel ECG-based methodology for detecting hypertensive patients. It shows that it is easily implementable and capable of classifying with high accuracy.

Sentiment Analysis Based on User Charging Comments

Mincong Tang

Xuzhou University of Technology, China

tang12290@gmail.com

In order to better understand the charging needs of electric vehicle users, this paper mainly conducts sentiment analysis based on comments data from electric vehicle users about charging stations to provide appropriate suggestions for improving charging infrastructure. The main research work carried out in this paper is as follows: (1) Crawl related comments from users about charging piles and other charging infrastructure and pre-process them using Jieba word segmentation. (2) The pre-processed comment data is sentiment analysed using the SnowNLP library, calculating sentiment scores and performing regional and temporal analysis based on the sentiment scores. (3) The processed comment data is analysed using the LDA topic model, focusing on the determination of user attention topics and changes in topic words. (4) Combining the results of sentiment analysis and topic analysis to propose targeted suggestions.

Comparative Analysis of Clinical Decision Support Systems for Eye Fundus Images

Steponas Tolomanovas, Jolita Bernatavičienė, Povilas Treigys

Institute of Data Science and Digital Technologies
Vilnius University

steponas.tolomanovas@mif.vu.lt

In ophthalmology there is a crucial need for early and accurate identification of eye diseases, particularly in areas with limited resources or in rural locations. With the evolution of handheld fundus cameras in recent years, remote and small clinics can now capture high-quality retinal images, allowing them access to advanced eye care. These fundus cameras are being enhanced with the capabilities of artificial intelligence through clinical decision support systems, enabling timely detection of eye diseases. In this study, we conducted a comparative analysis of 6 clinical decision support systems for eye fundus images, aiming to gauge their versatility, accuracy, and integration across different operational settings by gathering and reviewing information available from the pages of each CDSS. Our aim was to gauge their versatility, accuracy, and integration across different operational settings. Our analysis encompassed several metrics: diagnostic accuracy, interoperability with different imaging systems and diseases, architecture, and openness as claimed by the providers. The results revealed that most systems do not provide diagnostic accuracy metrics on their pages, leaving a significant gap in understanding their performance. In addition, the analysis revealed that most systems focused on the detection of a narrow range of diseases, most commonly diabetic retinopathy, significantly limiting their utility for a wider range of eye conditions. Furthermore, several systems only supported a limited set of imaging systems, further limiting their applicability, particularly in diverse operational settings that employ a variety of imaging devices.

Financial Data Anomaly Detection Through Behavioral Change Indicators

Ilona Veitaitė[1], Audrius Lopata[2], Saulius Gudas[3]

[1] Institute of Social Sciences and Applied Informatics, Vilnius university
[2] Faculty of Informatics, Kaunas University of Technology
[3] Institute of Data Science and Digital Technologies, Vilnius university

ilona.veitaite@knf.vu.lt

The research explores a method for analyzing financial data through the usage of Behavioral Change Indicators (BCI). It presents the BCI-based method for identifying financial anomalies, alongside the design of a architecture for detecting these anomalies. Furthermore, it introduces a theoretical architectural approach for recognizing financial irregularities, which has been effectively implemented using the Camunda business rules engine and rigorously tested with real financial data. Key Performance Indicator (KPI) stated as a quantifiable measure of performance over time for a specific goal. KPIs provide objectives for teams and insights that help members across the organization make better decisions. Key Performance Indicators practically help every area of the business move forward at the strategic level. For managing the fiscal health of the organizations are required to know about the performance as well as about Behavioral Change Indicators (BCI). Financial data of the organization may be analyzed from different views for discovery of certain patterns and/or anomalies. In this research there is provided list of financial data set anomaly detection steps; there is described how to use Key Performance Indicators (KPI) in organization's finance management process. There is presented detailed example of financial data analysis using BCI. Provided BCI calculation and visualization example helps to define the benefits of BCI usage in financial data analysis. The presented results are part of deliverables of research project "Enterprise Financial Performance Data Analysis Tools Platform (AIFA)". The research project was funded by European Regional Development Fund according to the 2014–2020 Operational Programme for the European Union Funds' Investments under measure No. 01.2.1-LVPA-T-848 "Smart FDI".

Confidence and Prediction Intervals Usage in Maritime Traffic Awareness Evaluation Using LSTM Deep Neural Networks

Julius Venskus[1,2], Robertas Jurkus[1,2], Povilas Treigys[1]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Klaipeda University

julius.venskus@mif.vu.lt

Maritime traffic awareness stands at the crossroads of dynamic oceanic conditions, diverse vessel behaviours, and the intricacies of global trade pathways. The application of Long Short-Term Memory (LSTM) Autoencoder Deep Neural Networks has heralded significant advancements in the realm of maritime traffic prediction, offering a nuanced capability to capture temporal vessel movement patterns and anomalies. Yet, in the vast, unpredictable marine landscape, point predictions, no matter how accurate, can fall short without quantified uncertainties. This presentation ventures into the transformative integration of confidence and prediction intervals within LSTM Autoencoder models, enhancing maritime prediction reliability and robustness. Initially, we shed light on the unique challenges that maritime traffic prediction confronts, emphasising the nonlinearities and temporal dependencies of vessel movement data. The LSTM architecture, inherently suitable for sequence-based data, provides a foundational base for accurate trajectory predictions. Building on this, our advanced model not only captures and reconstructs vessel movement sequences but also estimates the associated uncertainties by generating confidence and prediction intervals. Automated Identification System (AIS) data collected in the Baltic Sea region is used for the evaluation, underscoring the model's prowess in offering a probabilistic view of vessel movements. By presenting both predicted trajectories and their corresponding intervals, stakeholders can assess the potential variability and risks in vessel paths, facilitating enhanced decision-making, from collision avoidance manoeuvres to port docking

priorities. Through real-world case studies, authors delineate instances where the integration of prediction intervals can dramatically shape operational and safety decisions, underscoring their indispensable value in maritime traffic awareness. In conclusion, our presentation shows the vital role of confidence and prediction intervals in LSTM neural network models, ushering in a paradigm shift towards uncertainty-aware, resilient maritime operations by evaluating coalition likelihood.

Evaluation of Consumer Confidence Indicators Using Social Media and Administrative Data

Akvilė Vitkauskaitė[1,2], Andrius Čiginas[1,2]

[1] State Data Agency (Statistics Lithuania)
[2] Vilnius University

akvile.vitkauskaite@stat.gov.lt

Consumer confidence is an important economic indicator assessed by a survey. The administration of the Consumer Confidence Index (CCI) lies under the purview of Eurostat, the statistical office of the European Union, with monthly calculations conducted across all member countries. The State Data Agency (Statistics Lithuania) is responsible for conducting the corresponding survey in Lithuania. The main objective of the consumer opinion statistical survey is to obtain information regarding consumers' intentions to make purchases, their saving capabilities, and their perceptions of the economic situation and its influence on their intentions. Social media platforms and administrative registers are alternative data sources that can improve the estimation. Social media platforms allow individuals to openly express their opinions and experiences, while administrative data can provide a comprehensive and objective source of information. This study examines the relationships between traditional survey-based indicators and consumer sentiment expressed on social media platforms. It investigates social media using data from X (Twitter) that is retrieved using the official Twitter API. We create a Social Media Indicator (SMI) using sentiment analysis on tweet text as auxiliary data for our time series analysis of CCI. This study also explores the potential of using administrative data, including key economic indicators like unemployment, inflation, and income data, as auxiliary variables to enhance the forecasting accuracy of CCIs. In general, obtaining data for research from popular social platforms such as Facebook and Instagram is challenging due to stringent privacy policies and data protection regulations. The data are easily available from X, but this platform is not

popular in Lithuania. Therefore, the representativeness of X data raises special issues in our study. Nevertheless, we aim to integrate the traditional survey data, social media sentiment, and essential economic data, while addressing challenges and uncertainties. We also examine the CCI nowcasting possibilities using the constructed SMI.

Customer Churn Prediction in the Software as a Service Industry

Eimantas Zaranka[1,2], Bohdan Zhyhun[1,2], Milita Songailaitė[1,2], Rūta Juozaitienė[1,2], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development

eimantas.zaranka@card-ai.eu

In the modern commercial environment characterised by a plethora of alternatives available to consumers for identical products, customer retention plays a pivotal role in sustainable business success. This research investigates customer churn prediction through the application of a diverse array of machine learning algorithms, including logistic regression, support vector machines, decision trees, random forests, and gradient-boosted trees. We use real-world data obtained from a company specialising in offering subscription-based services designed to enhance individuals' personal development. The dataset included business-related customer data such as money spent, the last payment date, total orders completed, and customer platform usage data, including the number of activities completed and the timeframe since account creation, etc. Several experiments were conducted, involving the exploration of various feature subsets obtained via "Boruta", "Boruta Shap", decision tree feature importance, and correlation coefficient techniques to identify the most promising feature set within different prediction time horizon windows. The trained models underwent evaluation based on multiple performance metrics, including accuracy, precision, recall, and F1 score. This investigation concluded that the gradient-boosted trees algorithm emerged as the most promising model for predicting customer churn, delivering an impressive overall accuracy of 95.5%.

Effects of Interface Layout Structure on Usability and Content Comprehension

Līga Zariņa[1], Jurģis Šķilters[1],
Solvita Umbraško[2], Santa Bartušēvica[1]

[1] Laboratory of Perception and Cognitive Systems, Faculty of Computing, University of Latvia
[2] Faculty of Education, Psychology and Art, Faculty of Computing, University of Latvia

liga.zarina@lu.lv

Although interacting with visual displays of websites is involved in learners' everyday online practices, little is known about the impact of geometric principles of interface layout on learning. We selected typical interface layout structures (e.g., columns, boxes, grid) from popular websites in Latvia used for different purposes (e.g., education, finances, entertainment) and created different website layout prototypes, varied based on symmetry features. In a quasi-experiment (n=56), the preference for prototypes was rated according to aesthetic and usability criteria (Lavie and Tractinsky, 2004). Next, in an online experiment (n=60), we used two preferred educational website layouts to test text comprehension in STEM and humanities topics. The results show the effects of layout complexity, symmetry type, and symmetry axis orientation on interface usability.

Lavie, T., & Tractinsky, N. (2004). Assessing dimensions of perceived visual aesthetics of web sites. International journal of human-computer studies, 60(3), 269-298.

Examining the Link Between Intangible Cultural Heritage and Tourism Through Correlational Analysis

Monika Zdanavičiūtė[1,2], Marco Scholtz[3], Kaat De Ridder[3], Tomas Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development
[3] Thomas More University of Applied Sciences, Belgium

monika.zdanaviciute@card-ai.eu

The tourism industry is a multifaceted sector intricately linked not only to economic growth but also to the understanding of diverse cultures and cultural exchanges. One of the most popular forms of tourism is heritage tourism, specifically focusing on Intangible Cultural Heritage (ICH). Within ICH lies a wealth of experiences encompassing histories, traditions, festivals, religious rituals, and various other living expressions of culture. Unfortunately, the potential of ICH in the tourism sector is often undervalued and underutilized. This is partly due to the fact that the relationship between ICH and tourism has not been extensively analyzed. The objective of our study was to analyze the connection between ICH and tourism based on official data. The analysis utilized the UNESCO ICH dataset containing information from 128 countries about 562 different ICH elements spanning from 2008 to 2020. Additionally, data from the United Nations World Tourism Organization, comprising arrival statistics from 218 countries between 1995 and 2020, was employed. Pearson's correlation coefficient was applied to examine the relationship between the number of ICH elements and tourist arrivals, the relationship between specific ICH element numbers and tourist arrivals, and the relationship between the number of ICH elements and changes in tourist arrivals across various time intervals. Upon conducting the analysis, it becomes evident that countries with a higher number of ICH elements experience a greater increase in tourist numbers.

Development and Evaluation of a Tool for the Retrieval of Semantically Related Words for Lithuanian Language

Bohdan Zhyhun[1,2], Justina Mandravickaite[1,2]

[1] Vytautas Magnus University
[2] Centre for Applied Research and Development

bohdan.zhyhun@vdu.lt

In this work, we developed a tool for retrieving nearest word approximations using word embeddings for the Lithuanian language. Understanding the problem of the lack of ready-to-use, user-friendly tools for less studied languages, our solution includes a Graphical User Interface (GUI) and a request-response mode, which allows easy interaction and investigation of semantically related words for different purposes. We experimented with the following word embedding models: FastText with pre-trained embeddings; improved FastText model by re-training our corpus on top of its ready-made model; Word2Vec developed from scratch; GloVe model using pre-existing embeddings. We used Pedagogic Corpus of Lithuanian (Kovalevskaitė & Rimkutė, 2020) for training and re-training. It consists of written data and orthographically transcribed spoken data and it has 669 000 words in total. Our comparison of these models gives insights into their performance and suitability for the Lithuanian language. Moreover, the tool's design and capabilities are designed for a wide range of applications, from linguistic research to technology development projects that require linguistic precision and depth. The importance of our tool underscores the pressing need for robust linguistic tools catered to commonly less studied languages. It serves as a bridge in reducing the technology gap and offers a promising direction for further advancements in natural language processing for the Lithuanian language.

14th Conference

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Compiler **Jolita Bernatavičienė**
Prepared for press and published by
Vilnius University
Institute of Data Science and Digital Technologies
4 Akademijos St., LT-08412 Vilnius

Vilnius University Press
9 Saulėtekio Av., III Building, LT-10222 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt
Books online bookshop.vu.lt
Scholarly journals journals.vu.lt