12th Conference on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Druskininkai, Lithuania, Hotel “Europa Royale”
http://www.mii.lt/DAMSS

December 2–4, 2021

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

12th Conference on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

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

December 2–4, 2021

VILNIUS UNIVERSITY PRESS
Vilnius, 2021

Co-Chairmen:

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

Programme Committee:

Prof. Juris Borzov (Latvia)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Janusz Kacprzyk (Poland)
Prof. Ignacy Kaliszewski (Poland)
Prof. Yuriy Kharin (Belarus)
Prof. Tomas Krilavičius (Lithuania)
Prof. Julius Žilinskas (Lithuania)

Organizing Committee:

Dr. Jolita Bernatavičienė
Dr. Olga Kurasova
Dr. Viktor Medvedev
Dr. Martynas Sabaliauskas
Laima Paliulionienė

Contacts:

Dr. Jolita Bernatavičienė
jolita.bernataviciene@mif.vu.lt
Dr. Olga Kurasova
olga.kurasova@mif.vu.lt
Tel. +370 5 2109 315

https://doi.org/10.15388/DAMSS.12.2021
ISBN 978-609-07-0673-2 (print)
ISBN 978-609-07-0674-9 (digital PDF)

Preface

DAMSS-2021 is the 12th international conference on data analysis methods for software systems, organized in Druskininkai, Lithuania. The same place and the same time every year. The exception was 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 is back on track. When we organize it, the main priority is the health of the participants. There is a possibility to read the oral presentations online. Plenary sessions can be attended remotely, even if the conference participant is staying in a hotel. Efforts will also be made to regulate the flow of participants. But this conference is not a hybrid conference, as the main objective is live interaction between researchers.

History of the conference starts from 2009 with 16 presentations. It started as the workshop and has now grown into a well-known conference. The idea of such workshop came up at the Institute of Mathematics and Informatics that now is the Institute of Data Science and Digital Technologies of Vilnius University. The Lithuanian Academy of Sciences and the Lithuanian Computer Society supported this idea. This idea got approval both in the Lithuanian research community and abroad. The number of this year presentations is 63. The number of registered participants is 92 from 7 countries. That's not much less than in 2019. Of course, the number of participants coming from abroad is much lower, as there are still doubts about the safety of long-distance travel. The conference brings together researchers from six Lithuanian universities. This makes the conference the main annual meeting point for Lithuanian computer scientists.

The main goal of the conference is to introduce the research undertaken at Lithuanian and foreign universities in the fields of data science and software engineering. Annual organization of the conference allows the fast interchanging of new ideas among the research community. Four IT companies supported the conference this year. This means that the topics of the conference are actual for business, too. Topics of the conference cover Artificial Intelligence, Big Data, Bioinformatics, Blockchain Technologies, Business Rules Software Engineering, Data Science, Deep Learning, Digital Technologies, High-Performance Computing, Machine Learning, Medical Informatics, Modelling Educational Data, Ontological Engineering, Optimization in Data Science, Signal Processing, Visualization Methods for Multidimensional Data.

This book gives an overview of all presentations of DAMSS-2021.

Partner

International Federation for Information Processing
ifip.org

Supported by:

General sponsors

VTeX
vtex.lt

Novian Systems
novian.lt

Main sponsor

Asseco Lithuania
asseco.lt

Sponsor

Baltic Amadeus
www.baltic-amadeus.lt

Power Analysis of Multivariate Goodness of Fit Tests

Jurgita Arnastauskaitė, Tomas Ruzgas, Mindaugas Bražėnas

Kaunas University of Technology
jurgita.arnastauskaite@ktu.lt

In modern data analytics, decisions making involves hypotheses testing. It is a common practice to check the assumption of data normality. Which dictates the choice of data analysis methods (parametric or non-parametric). The assumption of normality can be checked graphically, but a more consistent option is to test the goodness of fit hypothesis. Despite the fact that a lot of statistical test have been developed since the 20th century, analysis of multivariate data remains challenging. The purpose of this study is to perform a power analysis of multivariate goodness of fit hypothesis test for the assumption of normality for different data sets and to compare the results obtained with our proposed test. Thus, we proposed a new powerful multivariate test (MIDE), which is based on the mean absolute deviation of the empirical distribution density from the theoretical distribution density. In this test, the density estimate is derived by using a inversion formula. To show advantages of our test an exhaustive comparative study of multivariate tests was performed. For this purpose, a lot of multivariate data sets of non-normal distributions were generated. For the comparison, the power of well-known test and our test was evaluated empirically. Based on the obtained modelling results, it can be concluded that the MIDE test.

Mental Health Effects of Covid-19 Lockdown: Empirical Findings and Multivariate Regression Model

Stefano Bonnini, Michela Borghesi

University of Ferrara, Italy
stefano.bonnini@unife.it

The Covid-19 pandemic has quickly turned into a psychological emergency: it has in fact changed the lives of individuals. The governments of various countries, to limit the spread of the virus, have introduced containment measures, which on the one hand have served to contain it and limit its spread, on the other have affected people's mental health. This work investigates the possible effects on mental health of the covid-19 lockdown, by analyzing the data of a survey carried out by the University of Milano Bicocca entitled "Forced social isolation and mental health: A study on 1006 Italians under COVID-19 lockdown". Specifically, the data collected in the survey have been analyzed in order to provide new original findings about the topic. In particular, given the multidimensional nature of "mental health", a multivariate regression model has been defined to study the relationship between mental health and some specific factors that may have worsened psychological status of people under the lockdown. A non-exhaustive list of these factors includes space adequacy, age, occupation, educational level, number of social contacts, perceived closeness of online and offline contacts. The work represents an output of the project "Identification of socio-demographic and biological causes of the high mortality from COVID-19 in the elderly population (ICE COVID-19)" funded with the "5x1000" contribution of the year 2018 by University of Ferrara.

Forbearance Prediction Using XGBoost and Light GBM models

Dalia Breskuvienė

Danske Bank A/S Lithuania Branch
dalia.breskuviene@gmail.com

Bank carefully tracks its weak and vulnerable (W&V) clients segment. One of the ways to become a W&V customer is when a facility becomes forborne. It is essential to follow the "as is" situation and look into the future of W&V segment development. The model aims to predict if the customer's facilities will become forborne during the upcoming six months. The data cleaning part requires a trade-off between the amount of source data and functionality, as data granularity is on the customer facility level each month when the scope of the data is 24 months. The KMeans clustering model in PySpark was used to group facilities by some of their characteristics to reduce the number of rows in the data source. The data source was imbalanced by the number of forbearance cases in each cluster.

The XGB and Light GBM Classification models were chosen for predicting future forborne facilities as they recently showed the best performance on the bank data in other problem-solving tasks. XGBoost and Light GBM is high-performance gradient boosting frameworks based on a decision tree algorithm. Light GBM differs from other decision tree algorithms because it splits the tree leaf-wise instead of the tree depth-wise or level-wise.

The features selection algorithms and consultations with advisors and other credit risk-related specialists helped to understand which features could identify changing patterns in the credit portfolio.
A complete model would allow risk managers to take strategic actions based on the predictions.

User Behaviour Analysis Based on Similarity Measures to Detect Anomalies

Arnoldas Budžys, Viktor Medvedev, Olga Kurasova

Institute of Data Science and Digital Technologies
Vilnius University
arnoldas.budzys@mif.stud.vu.lt

User behaviour analysis is based on statistical methods and machine learning to detect significant deviations from users' standard (ordinary) patterns or trends. These behavioural analysis results are useful in the context of cybersecurity as they allow critical infrastructure managers and administrators to anticipate potential security incidents. The user identification approach, based on keyboard keystroke dynamics, provides a secure interface between the user and the environment (e.g. computer, web environment, server) throughout the session. A typical architecture of a keystroke dynamics authentication system consists of data collection, feature extraction, classification and result evaluation. During the monitoring process, by continuously comparing the user-generated keyboard signals with the previous signals, anomalies can be detected and cybersecurity experts can be alerted to possible identity breaches. To solve this problem, similarity measures (DTW, correlation coefficient, extended Frobenius norm, etc.), one-class naïve Bayes, support vector machine can be used. The preliminary results for anomaly detection are presented in this paper after processing the test data using similarity measures.

Machine Learning vs. Fuzzy Inference Methods for Predicting the Oil Spill Consequences with Small Data Sets

Anastasiya Burmakova, Diana Kalibatienė

Vilnius Gediminas Technical University
anastasiya.burmakova@vilniustech.lt

Big data is usually needed for machine learning and fuzzy inference methods to predict consequences in different application areas. However, in practice, learning, inferencing, and prediction from small data remain a key challenge in machine learning and fuzzy inference systems. No exception is the prediction of the oil spill consequences on the ground environment, for which only a limited data set is available. In this study, we have used several machine learning methods (support vector regression (SVR), Decision trees, Ensembles, and Gaussian Progress Regression) and the adaptive neural fuzzy inference system (ANFIS) to predict the oil spill consequences on the ground environment from small data sets of real oil spill objects. An additional pre-training of methods was performed with the synthetic data obtained from the oil spill prediction mathematical model. Results obtained during the experiments have shown that ANFIS compared with the machine learning methods, exhibits higher prediction accuracy and better performance. The ANFIS method comparing to the machine learning method (i.e., Gaussian Progress Regression) have obtained the correlation coefficient (R = 0.9992), the coefficient of determination (R2 = 0.9984) and the root-mean-square error (RMSE = 0.001%).

An Architecture of Cognitive Health Evaluation and Monitoring System

Eglė Butkevičiūtė, Liepa Bikulčienė,
Tomas Blažauskas

Kaunas University of Technology
egle.butkeviciute@ktu.lt

Together with physical health condition monitoring, professional athletes focus on how cognitive-mental abilities like reaction times, anticipation, risk taking, etc. influence their performance in competitions. Cognitive health differences are visible when comparing high-performance athletes to novices – certain mental abilities like decision making or anticipation are much more developed in elite athletes. Registration of physiological parameters together with decision making and cognitive tests is an important part in self-regulation programs of athletes. The main purpose of this work is to create a system prototype that uses virtual reality devices that support WebVR. The implementation of AI techniques gives feedback and identifies factors that are bottlenecking cognitive health performance. The improvements in smart coaching of professional athletes may improve their mental abilities. This work presents a recent Kaunas University of Technology research in the ITEA-2019-19008 Inno4Health project "Stimulate Continuous Monitoring in Personal and Physical health".

Newest Machine Learning Password Guessing Techniques

Andrius Chaževskas

Institute of Data Science and Digital Technologies
Vilnius University
andrius.chazevskas@mif.stud.vu.lt

Password guessing is essential for forensic encrypted data examination since the data must be decrypted first. The most common password guessing attacks are dictionary and brute-force. A brute force attack is an attack when all possible passwords from a defined set of characters are tested until the correct one is found. A dictionary attack is based on trying all the words in a pre-arranged dictionary. The main drawback of the brute-force attack is the size of a set of all possible password candidates, which grows exponentially with the length of the password. The ability of laboratories to rely on brute force attacks depends on available hardware resources (how many guesses attempts can be made per second) and is limited. The analysis of leaked password databases shows that users tend to use easy-to-remember passwords. It means that the passwords usually exhibit a logical structure; they are not just random character sets. Modern automated password guessing strategies relying on machine learning and natural language processing try to exploit this defect. This topic presents a survey of the latest password guessing methods and strategies applied in password guessing techniques.

Acoustic Analysis for Vocal Fold Assessment – Challenges, Trends, and Opportunities

Monika Danilovaitė, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University
monika.danilovaite@mif.stud.vu.lt

Between 3 % - 9 % of USA (United States of America) population is affected by voice quality disorders but only a small part seek treatment. Researchers try to find solution to this problem with various proposed methods and tools for vocal folds assessment. However, scope of available methods is broad, and it is difficult to critically evaluate research trends and developments. The aim of this work is to review trends of vocal folds assessment. Systematic mapping study method was used to classify and assess quantitatively selected research documents. Data was inferred, results show that vocal fold's assessment is influenced by general computer science trends. A substantial proportion of the studies use machine learning methods (51% of selected studies used at least one method). Feature based analysis predominates research. Authors publish high classification accuracy results under controlled environment. Lack of studies where change of vocal folds state is assessed was observed. General quantitative and objective indicators would allow assessment of vocal folds state, state change and indicators' link with pathologies.

Software for Automatic Anonymisation of Radiographic Images

Rugilė Dauliūtė[1], Eimantas Mačius[1],
Gediminas Danys[2], Justas Trinkūnas[2,3],
Roma Puronaitė[2,4,5], Rolandas Bėrontas[2]

[1] Innovation incubator "KTU Startup Space" of Kaunas Technology University
[2] Vilnius University Hospital Santaros Klinikos
[3] Vilnius Gediminas Technical University
[4] Institute of Data Science and Digital Technologies, Vilnius University
[5] Faculty of Medicine, Vilnius University
roma.puronaite@santa.lt

With the entry into force of The General Data Protection Regulation (GDPR), there is a strong focus on the anonymisation of medical images and the protection of patient data. It is a crucial task to maintain the quality of the image and its suitability for processing. The main goal is to quickly and qualitatively depersonalise images and thus open images to science and business, enabling new algorithms and innovative technologies development. We have developed a fully automated program for radiographic picture anonymisation. The program can automatically analyse the X-ray images or radiographs. Medical images are called DICOM. They contain not only picture data, but also metadata, which needs to be hidden. One of the biggest challenges – sensitive information embedded on the picture pixels. Our software automatically detects text using machine learning algorithms and covers it. Every image has generic metadata which contains sensitive information about the patient and hospital. Our software outputs the result files in a convenient data structure for easy user experience. Doctors can easily select pictures for anonymisation and the program does the task in several minutes – depending on the number of images. 1 picture takes about 5s to anonymise. Previously it took several minutes to anonymise only one picture manually. The program is compatible with images acquired from all X-ray machine providers, including 30 of Santaros clinics' X-ray machines. Our software can recognise them and safely transform and cover sensitive information. The anonymisation is done sufficiently to correspond to current

DICOM standards. The main benefit is that end users are saving time on manual anonymisation of every image. Startups and scientists are benefiting from the fact that they are eligible to take images and use them in their work, for example producing machine learning algorithms. After anonymisation – the actual image does not lose quality, so it could be further analysed. We plan to improve the anonymisation algorithm and adapt the software to other types of radiological images, including multi-frame images.

On the Parallelization of the Geometric Multidimensional Scaling

Gintautas Dzemyda, Viktor Medvedev,
Martynas Sabaliauskas

Institute of Data Science and Digital Technologies
Vilnius University
viktor.medvedev@mif.vu.lt

A well-known procedure for mapping data from a high-dimensional space to a lower-dimensional space is Multidimensional Scaling (MDS). This algorithm keeps the distances in the low-dimensional space as close as possible to the distances in the original multidimensional space. Although MDS demonstrates great versatility, it is computationally demanding. Conventional approaches to the MDS method are limited when analyzing very large datasets, as they require very long computation times and large amounts of memory. The MDS method is used in many data analysis applications. However, despite a number of studies in which MDS methods have been applied to data mining, the number of data points to be analyzed has been limited by the high computational complexity of MDS. A new Geometric MDS method with lower computational complexity has been developed (G. Dzemyda and M. Sabaliauskas. Geometric multidimensional scaling: A new approach for data dimensionality reduction. *Appl. Math. Comput*. 409, 125561 (2021). https://doi.org/10.1016/j.amc.2020.125561; G. Dzemyda and M. Sabaliauskas. 2021. New capabilities of the geometric multidimensional scaling. In *Trends and Applications in Information Systems and Technologies*. WorldCIST 2021, A. Rocha et al. (Ed.). *Advances in Intelligent Systems and Computing*, Vol. 1366. Springer, 264–273. https://doi.org/10.1007/978-3-030-72651-5_26), making MDS more applicable to large-scale datasets. The results allow us to extend the application of this method to new and efficient ways of visualizing large-scale multidimensional data and reducing its dimensionality. A way to minimize MDS stress has been developed using the ideas of Geometric MDS, where all points in a low-dimensional space change their coordinates simultaneously and independently dur-

ing a single iteration of stress minimization. Its efficiency depends on the choice of parallelization strategy associated with different ways of grouping the analyzed data, simultaneous computation of multiple coordinates in low-dimensional space, and the number of multi-core processors used for the computation. The proposed Geometric MDS allows the implementation of parallel computing for the dimensionality reduction process of large-scale data using multithreaded multi-core processors or parallel coprocessors such as GPUs. We examine how the computational speed of multidimensional data visualization varies depending on the strategy chosen and the number of processors.

This research has received funding from the Research Council of Lithuania (LMTLT), agreement No S-MIP-20-19.

TCEM: An Exhaustive Text-to-Code Evaluation Metric for the Auto-Generated Code Snippets

Domnica Dzitac[1], Wajahat Mirza[1],
Bokang Jia[2], Simona Dzitac[3]

[1] New York University Abu Dhabi, UAE
[2] Mohamed bin Zayed University of Artificial Intelligence, UAE
[3] Department of Energy Engineering, University of Oradea, Romania
domnica.dzitac@nyu.edu

Translating natural language to code snippets (text-to-code) is a task that has been repeatedly attempted using several different models in recent years. However, no one investigated how GPT-3, one of the newest and most powerful transformers performs on this task. In this research paper, we contribute to related literature (1) by showing how GPT-3 performs on the text-to-code task and (2) by creating TCEM, a more powerful and accurate evaluation metric than BLEU score. The latter contribution, TCEM, covers edge cases, evaluates the functionality of the code on each line level, and provides overall cumulative mean score percentage based on the successfully passed unit tests. This way, we provide a robust metric of evaluation for the text-to-code tasks that focuses on the functionality of the generated code snippets rather than the semantics and words they contain.

Application of MCDM Techniques for Consensus Protocol Selection

Ernestas Filatovas, Remigijus Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University
ernestas.filatovas@mif.vu.lt

Blockchain technologies have already engaged great attention from industry and academia. A consensus protocol plays a key role in the complex architecture of a blockchain system. It ensures that all participants (nodes) of a blockchain network agree on the transactions without a central authority. A wide range of blockchain consensus protocols have been developed considering different aspects (e.g., energy consumption, scalability, latency, throughput, fault tolerance, etc.) and types of systems (private, public, consortium). The requirements of blockchain systems can differ considerably. Therefore, selecting a suitable consensus protocol to meet the needs of a specific system is challenging. To facilitate the selection process, various Multi-Criteria Decision-Making (MCDM) techniques could be employed. Here we present an MCDM-based framework for identifying preferable consensuses for each type of blockchain system based on the specified criteria and their weights. We also demonstrate the potential of the framework on the popular blockchain consensus protocols.

Web Tools for Analysing Location-Based Data

Pasi Fränti

School of Computing, University of Eastern Finland
franti@cs.uef.fi

Tracking people location has become every day practice and created lots of new geotagged data. This presentation gives an overview for recent methods on analysing collected geotagged data via developed web applications. These include various GPS trajectories analysis including similarity, averaging, reduction, move type detection, distance calculation, clustering, and optimizing facility locations. Most methods are publicly available on the web either as demonstrations, APIs or web tools where user can upload his/her own data.

Numerical Simulations of Heat and Mass Exchange Between Human Skin and Textile Structures

Aušra Gadeikytė, Donatas Sandonavičius, Vidmantas Rimavičius, Rimantas Barauskas

Department of Applied Informatics
Kaunas University of Technology
ausra.gadeikyte@ktu.lt

The main function of clothing is to protect the human body from hazardous environments. Nowadays, the three-dimensional textile is used as a moisture and thermal regulating layer in multi-layer textile packages (e.g. protecting clothing, outdoor clothing), medical bandages. One of the challenges for clothing designers is to ensure thermal comfort between human skin and fabrics. According to the literature, the most important properties of thermal comfort are air permeability, water-vapor resistance, and thermal resistance. Modern finite element computing technologies allow for a highly realistic representation of the physical processes and can even be used to replace experiments. In this work, we present computational techniques and finite element models that allow predicting air permeability, water-vapor resistance, thermal resistance coefficients on a micro-scale. The models can be applied in the development of passive and active cooling systems. The numerical simulations were performed using Comsol Multiphysics and Matlab software.

Application of CNNs for Brain MRI Image Segmentation

Rokas Gipiškis, Olga Kurasova

Institute of Data Science and Digital Technologies
Vilnius University
rokas.gipiskis@mif.vu.lt

Semantic segmentation of brain tumors based on magnetic resonance imaging (MRI) is a crucial step in determining their type and location, as well as in contributing to a more accurate diagnosis for a patient. We evaluate the application of convolutional neural networks (CNNs) for segmenting gliomas, a type of primary brain tumors, in different anatomical planes in terms of Sorensen-Dice similarity coefficient. We identify promising encoder-decoder networks (2D U-Net, Attention U-Net, SegNet together with 3D implementations) and their modifications and present a primer on model-agnostic post hoc explainability methods which can improve our model's output. Explainable and interpretable models are important in building trust in automated decision-making systems, especially in high-impact areas. Medical experts can assess if a given explanation corresponds to relevant diagnostic criteria. We also discuss difficulties associated with formalizing the notion of interpretability and evaluating it within a medical domain.

Intelligent Data Capture in Digitized Business Documents

Rolandas Gricius, Igoris Belovas

Institute of Data Science and Digital Technologies
Vilnius University
rolandas.gricius@mif.stud.vu.lt

Today most documents are produced in digital format directly, removing the need for OCR. Unfortunately, documents are primarily in free form. Thus data and information still need to be extracted for further processing in information systems. We aim to present possible approaches in dealing with this problem.

Causal Interactions of the Agile Activities in Application Development Management

Saulius Gudas, Karolis Noreika

Institute of Data Science and Digital Technologies
Vilnius University
saulius.gudas@mif.vu.lt

Agile management methods and tools are being widely used to manage Enterprise Application Software (EAS) development. However only every third EAS development project is successful in terms of time, cost, and scope constraints, and users are satisfied with it. Our experience using Agile management tools like Atlassian "Jira" shows the lack of coordination between software development and business management activities. We focus on the modelling of causal interactions of the Agile activities such as Theme – Initiative – Epic – User story for bridging business strategy and application software development. The causal modelling paradigm is used to construct an internal model of coordination interactions and taxonomy of coordination types. The semantics of management interactions is based on the information transmitted between Agile items. There are two coordination types of the Agile activities hierarchy – vertical and horizontal coordination. The content of Agile activities is conceptualized as the management transaction (MT), which specifies the internal causation of activities.

The same principles of internal modelling are used in the Grey Systems Theory to reflect the perceived real-world information. The management transaction concept in our approach represents the causal model (grey model) of interactions in the Agile management hierarchy. The content of feedback between two adjacent levels such as Theme – Initiative, Initiative – Epic, Epic – User story is revealed using the concept of management transaction (MT). Themes and Initiatives are in the field of business management competency, and lower level (epics, user stores) in the field of software development project management. By defining the internal model of the Agile management hierarchy using the MT construct, the obtained causal knowledge is expressed as a new

attribute „capability“ in an Agile management tool like „Jira“. „Capability“ is considered here as a strategy-based functional requirements. This new attribute implements top-down strategic objectives deployment in EAS system development at all Agile management hierarchy levels. This enhancement to Agile’s management tools ensures that the integrity of the program’s project content is automatically checked against the company’s strategic objectives. This helps to reduce the mismatch between business strategy execution and software development management activities and ensures that the effort for EAS development is dedicated to work that contributes to the strategic objectives of the organization.

Estimating Energy Consumption of Ethereum Network

Aleksandr Igumenov, Ernestas Filatovas,
Viktor Medvedev, Remigijus Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University
aleksandr.igumenov@mif.vu.lt

Since the advent of blockchain technology, its various applications have grown rapidly. However, the use of blockchain has also revealed many challenges. One of them is the ever-increasing consumption of electricity used for mining. For example, energy usage for mining Bitcoin is comparable to some countries. Recently, several approaches and methods have been proposed to estimate the energy consumption of blockchain networks. The development of different methodologies is related to the problems of planning and distribution of electricity in the market and the environmental situation in general. With knowledge of the electrical energy consumption in this sector, governments can react dynamically to the situation by adopting laws affecting environmental improvements. However, most of the studies and proposed methodologies focus on calculating the Bitcoin network's energy consumption and pay insufficient attention to other large blockchain networks as Ethereum is. This study reviews existing state-of-the-art methods for estimating the energy consumption of the Ethereum network.

Investigation of Abnormal Prostate Region Detection Using Different Modality Combinations of mpMRI Scans

Justinas Jucevičius[1], Povilas Treigys[1], Jolita Bernatavičienė[1], Mantas Trakymas[2], Ieva Naruševičiūtė[2], Rūta Briedienė[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] National Cancer Institute
justinas.jucevicius@mif.vu.lt

There are many prevention programs in effect for various organ cancer nowadays and prostate cancer is not an exception. Prostate cancer is not only the second most frequent tumor among men, but is also one of the most morbid tumors worldwide. Lithuania has adopted a law for funding a program for early prostate cancer diagnosis on a national level in 2005. Despite biopsy being the only way to conclude a definite diagnosis of prostate cancer, it still misses up to 30% of clinically significant cancer and reason for that is taking samples from wrong location. National Comprehensive Cancer Center recommends using multiparametric magnetic resonance imaging (mpMRI) for aiding the diagnosis of prostate cancer by determining the location to perform biopsy on. According to the latest guidelines, radiologists must find abnormalities in at least three different mpMRI modalities for the region to become a biopsy candidate. The detection of these areas usually includes manual work, which depends on the experience of the personnel. In order to reduce the room for mistakes, a software is needed to aid with this task. This work is therefore dedicated to investigate the use of deep learning techniques for identifying regions to perform biopsy on and the effect of combinations of different modalities on segmentation results.

Unobserved Heterogeneity May Induce Ghost Triadic Effects in Relational Event Models

Rūta Juozaitienė[1,2], Ernst Wit[3]

[1] Vilnius University
[2] Vytautas Magnus University
[3] Università della Svizzera Italiana
ruta.juozaitiene@vdu.lt

Temporal network data often encode time-stamped interaction events between senders and receivers, such as co-authoring a scientific article or sending an email. A number of relational event frameworks have been proposed to address specific issues raised by modelling time-stamped data with complex temporal and spatial dependencies. These models attempt to quantify how individuals' behaviour, external factors and interaction with other individuals change the network structure over time. It is often of interest to determine whether changes in the network can be attributed to endogenous mechanisms reflecting natural relational tendencies, such as reciprocity or triadic effects. The propensity to form (or receive) ties can also be related to individual actors' attributes. Nodal heterogeneity in the network is often modelled by including actor-specific or dyadic covariates, such as age, gender, shared neighbourhood, etc. However, capturing personality traits such as popularity or expansiveness is difficult, if not impossible. A failure to account for unobserved heterogeneity may confound the substantive effect of key variables of interest. This research shows how node level popularity in terms of sender and receiver effects may mask ghost triadic effects. These results suggest that unobserved heterogeneity plays a substantial role in REM estimation procedure and influences the conclusions drawn from real-world networks.

Prediction of Vessels Trajectory Using Different Coordinate Systems

Robertas Jurkus, Povilas Treigys, Julius Venskus

Institute of Data Science and Digital Technologies
Vilnius University
robertas.jurkus@mif.stud.vu.lt

The global marine insurance report announces that the growing risk of major events (both human and non-human) remains among the ongoing challenges in maritime transport. One of the factors is vessel collisions and anomalies at the sea. Large historical data from automatic identification systems (AIS) are analysed to solve the problem of ship trajectory prediction. The most common attempt to improve accuracy is by evaluating the historical ship behaviour, learning the patterns and similarities of the predicted vessel movements. However, this paper shows that a better prognosis also may be reached by chosen a different trajectory calculation strategy. Typically, such features are the values of the polar coordinate system, so it has been proposed to transform the coordinates into two-dimensional space. The positioning of the vessels motion transformations were tested: coordinates transformation into a Cartesian system using Universal Transverse Mercator (UTM) projection. This case allowed to reduce the forecast error of almost 30% in the available data sample (it is almost 300 meters) by using the autoencoder architecture, compared to the longitude and latitude predictions. Overall, the research compares three recurrent network architectures (including their hyperparameter - cell size changes): bidirectional Long Short-Term Memory, autoencoder and gated recurrent unit networks. The model is applied to a real AIS historical dataset of the cargo vessel type trajectories in the Netherlands (North Sea) coastal region.

Features of the Python Active Lessons teaching methodology

Eimutis Karčiauskas

Kaunas University of Technology
eimutis.karciauskas@ktu.lt

As the number of data analysis tasks increases, the training of new researchers is becoming increasingly important. It is widely acknowledged that Python programming language is the most popular among the research tools. In order to get students interested in Python programming an active learning methodology has been developed. Its main features are:

- starting learning process with the construction of functions;
- representation of all algorithms by functions;
- as much knowledge as possible about the algorithms is presented in the source code;
- for each topic demonstrative functions are provided;
- the specified functions are provided for students to use in experiments;
- students receive the related tasks directly in the source code;
- the study tasks are relevant to real-world data.

This methodology was developed based on the programming and data structures teaching experience at the KTU Faculty of Informatics. This allowed many tasks to migrate to the school course adapting them to school curricula. Experience has shown that there is a significant increase in the effectiveness of training, as measured by the number of tasks completed. The understanding of the study material is also enhanced by the activity of students, as they do not have to waste time on data entry as it is in the classical teaching style, where the concept of functions is taught only as late as 7th - 8th topic.

Monitoring of Cancer Patients Using Smartphones Sensor Data

Gabrielė Kasputytė[1,2], Akvilė Landaitė[1,2], Rūta Juozaitienė[1,2], Adomas Bunevičius[3], Romas Bunevičius[3], Šarūnas Bagdonas[3], Jonas Venius[4,5], Tomas Krilavičius[1,2]

[1] CARD – Centre for Applied Research and Development
[2] Vytautas Magnus University
[3] ProIT, JSC
[4] Medical physics Department, National Cancer Institute
[5] Biomedical Physics Laboratory, National Cancer Institute
gabriele.kasputyte@card-ai.eu

The ubiquity of smartphones provides an opportunity to capture a passive collection of a wide range of behavioural data. In particular, smartphones are well-suited to objectively assess people's daily behaviours, such as physical movement, social interactions, and other activities. Furthermore, since passive data collection operates in the background without requiring any input from the users, it reflects their behaviour in the natural environment. Such data collection and analysis may be especially beneficial to the monitoring of patients between clinical encounters, as changes in these passively sensed digital biomarkers may reflect significant changes in functional status.

This research aims to optimize the monitoring of cancer patients by developing a model indicating active and passive periods, exertional activity during these periods and passive night-time period of each patient. The proposed approach analyses the variance of passively collected accelerometer data. Accelerometer variance exceeding predetermined thresholds indicates active and non-active periods. To evaluate the performance of the proposed framework, we use real-world data consisting of 11 de-identified patients' records.

Performance Benchmarking of Communication Patterns in Microservice Architecture

Justas Kazanavičius, Dalius Mažeika

Department of Information Systems
Vilnius Gediminas Technical University
justas.kazanavicius@vilniustech.lt

Microservice architecture is applied for cloud-native applications to decompose it into small functional units. A microservices-based application is a distributed low coupling system running on multiple processes or services, and therefore proper communication patterns between microservices must be defined. Communication has a significant impact on the application's performance and must be adapted depending on the application architecture, exchange data, deployment approach, and service topology. The study aimed to perform benchmarks of synchronous and asynchronous communication patterns between microservices and determine use cases for their application. The application-oriented criteria were introduced to evaluate communication patterns, and a microservice-based application was developed to perform experiments. Communication technologies based on remote procedure invocation and messaging protocols have been analyzed, and corresponding advantages and disadvantages have been identified.

Locations on Networks

Mindaugas Kepalas, Julius Žilinskas

Institute of Data Science and Digital Technologies
Vilnius University
mindaugas.kepalas@gmail.com

The general problem of locations on networks is presented: we want to open K facilities which are restricted to be on a network (it could be a road or a communication network, for example). This placement has some cost associated with it. The goal of the research is to develop tools suitable for solving the formulated problem: we seek to develop efficient algorithms which aim to find the minimal cost. We further illustrate that some real-life problems can be naturally formulated as locations on networks problem. Finally, we show a demo, illustrating the steps of the current algorithm on a geometrical problem.

Robust Statistical Analysis for BiCNAR(s) Time Series

Yuriy Kharin, Valeriy Voloshko

Research Institute for Applied Problems of Mathematics and Informatics
Belarusian State University
Minsk, Belarus
Kharin@bsu.by

A parametric family of parsimonious models for discrete time series $x_t \in A = \{0,1,...\}$, $t = 1, 2,...$, called Binomial conditionally nonlinear autoregressive time series of order s (BiCNAR(s)), is based on the property that all the conditional distributions of x_t under fixed s-prehistory $(x_{t-1}, ..., x_{t-s})$ are Binomial $Bi(N, \theta)$, where $\theta = \theta(x_{t-1}, ..., x_{t-s})$ depends on the prehistory in some special way [1]. Here we consider a distorted version of BiCNAR time series [1] contaminated with innovation outliers by ε – mixing with some fixed unknown outlier discrete distribution $u(x)$, $x \in A$ with fixed value of the expectation.

For the parameters of BiCNAR-model under innovation outliers we construct a new robust frequencies-based estimator (robust FBE) and prove its properties: consistency, asymptotic normality, efficiency, robustness. We also construct a consistent statistical estimator for the distortion level $\varepsilon \in (0,1)$. Algorithm for robust statistical forecasting of future states x_t, $t > T$, based on T previous observations $x_1, ..., x_T$ is proposed.

Theoretical results are illustrated by computer experiments on real financial data.

[1] Kharin, Yu., Voloshko, V. Robust estimation for Binomial conditionally nonlinear autoregressive time series based on multivariate conditional frequencies. J. Multivariate Analysis 185 (2021), Article 104777.

Examining the Self-Similarity Method for the Lombard Effect Recognition

Gražina Korvel[1], Krzysztof Kąkol[2],
Povilas Treigys[1], Bożena Kostek[3]

[1] Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania
[2] GPS Software, Gdansk, Poland
[3] Audio Acoustics Laboratory, Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland
grazina.korvel@mif.vu.lt

The Lombard speech is an effect discovered in 1909 by Etienne Lombard, a French otolaryngologist. Lombard speech is obtained when the speaker unconsciously increases their vocal effort in a noisy environment. Numerous studies on the Lombard effect have identified many features characteristic of this expression type. Well-known phenomena in the signal included in the Lombard speech are the following: the increased volume of the uttered speech, fundamental frequency rise, formant frequency rise, spectral tilt, duration of utterances (both elongation and shortening), prosody alteration. Most of these features can easily be determined, but observing changes in these features in the context of the Lombard speech is not so simple. The main reason for this is that the Lombard speech characteristics vary according to the noise level. In this research, the self-similarity method is employed for the Lombard effect recognition in the presence of noise. Self-similarity matrices based on acoustic parameters related to the Lombard effect are created and introduced as 2D space features at the CNN input.

The research carried out as a part of this work includes several stages: (1) recording the sound samples without and with Lombard effect (2) mixing the source files with white noise speech interfering signal (3) creating self-similarity matrices (4) constructing deep learning network (5) performing the statistical analysis of results. Exploratory, experimental results support the approach proposed.

On the Problem of Eigenvector Sign Ambiguity: Ad-Hoc Solution for Eigenvector Decomposition-Based Signal and Image Analysis

Algimantas Kriščiukaitis[1,2], Ana Rita Alves dos Santos Rodrigues[1], Robertas Petrolis[2], Vaidotas Marozas[1]

[1] Biomedical Engineering Institute, Kaunas University of Technology
[2] Lithuanian University of Health Sciences
algimantas.krisciukaitis@lsmuni.lt

Although eigenvalue decomposition based multivariate analysis methods as Principal Component Analysis or Singular Value Decomposition are well-established and can be performed using state-of-the-art algorithms, users still face the methods' inherent problem - eigenvector sign ambiguity. It can significantly impact the conclusions and interpretations drawn from the methods' results. Yet, no standardized mathematical method exists to resolve this problem, with only a few ad-hoc solutions published thus far. We have been facing this problem in two cases: (i) electrocardiosignal analysis: to determine the main spacial direction of depolarization and repolarization in the heart muscle. (ii) protein antibody array image analysis: to determine the weak chemiluminescence signal over background illumination. In both cases the direction of first eigenvector of covariation matrix is carrying essential information for further analysis. We propose an ad-hoc solution, based on intrinsic features of the original data. The method is based on the feature that its' distribution is always skewed to one predominant side. The electrical activity of the heart muscle during de- and re-polarization has a predominant polarity. Similarly, the chemiluminescence signal reflecting pixel values are always scattered from background illumination to the positive direction. Thus, we align the polarity of the first eigenvector coefficients with the polarity of the skewness of the corresponding original data. Testing this approach on more than 200 recordings of electrocardiosignals showed that only a few uncertainties remain in low amplitude signals with a biphasic shape. In 150 series of protein antibody array images, this approach showed no failure to determine the polarity. Both testings showed no errors in eigenvector calculations.

Evaluation of Artificial Intelligence Methods and Tools for Discovering Healthcare Data Patterns

Dalia Kriksciuniene[1], Virgilijus Sakalauskas[1],
Ivana Ognjanovic[2], Ramo Sandelij[2]

[1] Vilnius University
[2] University of Donja Gorica, Podgorica,Montenegro
dalia.kriksciuniene@knf.vu.lt

Together with the advancement of computer technology, the emerging paradox reveals that there is no shortage of data for analysis. The main problem is to decide which research method could be most relevant for finding a solution, what insights we can get from these data and what decisions can be proposed.

Aiming to increase the efficiency of health care and improve treatment methods by capturing patient information is often hampered by the poor quality of medical data collections, as in many cases, the healthcare data are unstructured and preserved in different systems and formats. There is no standard approach which methods of artificial intelligence and machine learning perform better in different problem areas and which computer tools could make their application more convenient and flexible. The research provides essential characteristics of methods traditionally applied in statistics and their advanced modifications, such as logit, probit models, K-means, and Neural networks. The performance of the methods, their analytical power and relevance to the healthcare application domain is illustrated by brief experimental computations for investigation of stroke patient database with the help of several readily available software tools, such as MS Power BI, Statistica, Matlab, Google BigQuery Machine Learning.

The discussion and comparative evaluation of the artificial intelligence approaches and the illustration of their performance by applying different AI methods and tools should help us to reveal the advantages of artificial intelligence and machine learning methods in the area of application of health data analysis in different cross-sections.

For this purpose, we will take a big real clinical record file and try to analyse it using various research methods. The database applied for the experimental research consists of the records of 944 different patients, characterised by 58 variables. This database is a collection of registered stroke cases of the neurology department of the Clinical Centre in Montenegro.

Artificial Intelligence Financial Distress Barometer (Companies' Case)

Dovilė Kuizinienė, Tomas Krilavičius

Vytautas Magnus University
d.kuiziniene@gmail.com

The purpose of the research is to create financial distress barometer for companies using artificial intelligence algorithms. The main novelty of this product is to look from traditional bankruptcy issues to a willingness to pay perspective. The willingness to pay perspective gives additional viewpoints, such as conscientiousness of paying its obligations on time, companies values for creditors, suppliers and customers, etc. For this reason, the analysis includes not only financial reports but additional data sources from the tax office, social security (e.g. SODRA), media, conditions from macroeconomic and industry perspectives. The analysis includes small and medium Lithuanian companies data from 2015 till 2021. Dynamic view, dimensionality, sample imbalance issues are considered in new financial distress barometer approach creation.

Application of Artificial Intelligence for Automatic Lending Decision Making Using Transactions Data

Dovilė Kuizinienė, Paulius Savickas, Tomas Krilavičius

Vytautas Magnus University
dovile.kuiziniene@card-ai.eu

Obtaining a loan for SMEs is often a complicated and delayed process that requires time and human resources from both the financial institution and the borrower. Automatization by incorporating machine learning techniques used will not only expedite the lending process but save costs and provide more affordable borrowing opportunities for SMEs.
Process automation requires additional data availability for less risky automatic decision-making. Therefore, the Payment Service Directive II (PSD II) will allow financial institutions to proceed with transaction data analyses for the borrowers, which is one of the key aspects of the real financial health of the company understanding. The PSD II implements an opportunity to better evaluate customer needs for the financial institution, with no relationship before. We are creating an automated model that incorporates transaction information that allows SMEs to quickly borrow money up to a certain amount of risk that is appropriate for financial institutions. After a fixed amount of lending money, where the automatic solution is too risky, this model should help the expert make decisions by granting larger loans to SMEs. For financial institutions, this model will save 70-80% of the expert time when assessing the ability of companies to borrow and thus process more loan applications.

Anomaly Detection in Systems Metrics

Rimantė Kunickaitė[1], Dovilė Servaitė[1],
Gabrielė Jenciūtė[1], Andrius Bumblauskas[2],
Miglė Bučelytė[2], Aldas Glemža[2], Tomas Krilavičius[1]

[1] Vytautas Magnus University
[2] Blue Bridge
dovileserv@gmail.com

The information technology (IT) sector is becoming an area where the highest value-added products are being developed. With the growing importance of IT systems, the public sector and business organizations are becoming responsible for data management and storage. More and more processes are managed by IT systems and the amount of data is constantly increasing, therefore security ensuring is becoming a challenge. Understanding the state of infrastructure and systems is essential for ensuring the reliability and stability of services. Information about the health and performance of the system helps to react to issues and make changes with confidence. One of the best ways to gain insight into the information of system status is to use a reliable monitoring system that collects metrics, visualizes data and alerts when things appear to be broken. In this research, metrics that represent the raw measurements of resource usage or behaviour that can be observed and collected throughout IT systems are analyzed. Anomaly detection in metrics is performed using ARIMA, TBATS, PROPHET, LSTM, and Isolation Forest methods, which can be applied to real-time operation to identify and report the detected anomaly as quickly as possible.

Competitive Facility Location Problems with Different Customer Behavior Rules

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

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] University of Murcia
algirdas.lancinskas@mif.vu.lt

Competitive facility location is important for firms which provide goods or services to customers and compete for the market share with other firms in a certain geographical area. There are various facility location models and strategies to solve them, which vary on their ingredients such as facility attraction function, decision variables, search space, objective functions, etc. Alongside the latter ingredient it is important to consider customer behavior rules, which describes customers choice on different facilities to buy a service. Our research is focused on a discrete competitive facility location problem for an entering firm which important for a new firm which enters the market considering competition for the market share with other facilities already in the market and different customer behavior rules. We are interested in investigation of impact of the customer behavior rule to the optimality of candidate locations and in finding robust solutions which are acceptable considering different customer behavior rules or in existence of uncertainty in customer behavior.

Vehicle Route Optimization Using Evolutionary Algorithms

Karolis Lašas[1,2], Arnas Matusevičius[1,2],
Rūta Juozaitienė[2], Tomas Krilavičius[2]

[1] CARD – Centre for Applied Research and Development
[2] Vytautas Magnus University
karolis.lasas@card-ai.eu

Every year growing population requires more sophisticated logistic systems to meet higher and higher demand. There are many subsystems to be enhanced when trying to improve the lacking behind logistic system. One such task is route optimization for a fleet of vehicles that performs a crucial role in the company`s daily routine. Most of the day-to-day operations are still facilitated by transportation dispatchers, who are responsible for routing the fleet in the most efficient manner. However, a constant stream of loads to handle, the number of available trucks and hard constraints that need to be satisfied throughout the delivery makes this problem extremely difficult. This research aims to develop a tool suggesting the most optimal routing solutions. In this project, we examine the performance of evolutionary algorithms, including genetic algorithm, ant colony optimization and tabu search method. We perform experiments using historical data on freight transportation in Europe. A directed graph represents freight transportation data where nodes indicate pick-up and delivery locations connected by edges. Therefore, the route is a subset of nodes where the first node indicates vehicle starting location followed by freights locations that need to be visited. The proposed framework should consider delivery time windows, loading times, cargo weight, possible allocation to terminals, and other related cargo parameters.

Natural Language Generation with Architecture of Transformers: A Case Study of Business Names Generation

Mantas Lukauskas[1,2], Tomas Rasymas[2], Domas Vaitmonas[2], Matas Minelga[2]

[1] Faculty of Mathematics and Natural Sciences
Kaunas University of Technology
[2] Zyro Inc.
mantas.lukauskas@ktu.edu

The continuous improvement of artificial intelligence/machine learning leads to an increasing search for the broader application of these technological solutions to structured and unstructured data. One of the applications for unstructured data is natural language processing (NLP). Natural language processing is the computer analysis and processing of natural language (which can be delivered and written) using various technologies. NLP aims at linguistically adapted various tasks or computer programs in human languages. Natural language processing is finding more and more different ways to adjust to real practical problems. These tasks can range from finding meaningful information in unstructured data (Pande and Merchant, 2018), analysing sentiments (Yang et al., 2020; Dang et al., 2020; Mishev et al., 2020), and translating the text into another language (Xia et al., 2019; Gheini et al., 2021) to fully automated human-level text creation (Wolf et al., 2019; Topal et al., 2021). This study aims to apply natural language modelling models and the architecture of transformers to generate high-quality business names. The dataset for this study consists of 350,928 observations/business names (299,964 training and 50,964 observations in the test sample). This data was collected using the websites of start-ups from all over the world. For different models comparison, the data set was divided into two parts. The training data set represented 80%, and the test data set 20%. The experiments in this study were performed using a Google Cloud Platform virtual machine with parameters:12 vCPUs, 78 GB random access mem-

ory (RAM), 1 x NVIDIA Tesla T4 GPU (16 GB VRAM). For the biggest models, the GPT-J-6B and GPT2-XL virtual machine parameters have been increased to 16vCPUs, 150GB of RAM, and 2x NVIDIA Tesla T4. Based on perplexity metrics, the best-rated model, in this case, is GPT. Meanwhile, considering only the new generation models, the best result is observed with the GPT2-Medium model. However, the results of the study show that people's assessment and assessment by perplexity are different. In human evaluation, it is observed that the best result is obtained using the GPT-Neo-1.3B model. The evaluation of this model is statistically significantly higher compared to other models ($p < 0.05$). Interestingly, the GPT-Neo-2.7B model has poorer results. Its evaluation does not differ statistically significantly from the GPT-Neo-125M model ($p > 0.05$), which is even 20 times smaller. A critical element in using the ZeRO3 optimizer is the high RAM usage. The highest RAM usage is observed in the most significant model GPT-J-6B. This usage is as high as 101 GB. It is also noted that GPT2-XL and GPTNeo-1.3B have a pretty similar RAM usage. The interesting fact is that the GPT model uses more RAM compared to GPT2 and DistilGPT2.

Analysis of Clustering Methods Performance Across Multiple Datasets

Mantas Lukauskas, Tomas Ruzgas

Faculty of Mathematics and Natural Sciences
Kaunas University of Technology
mantas.lukauskas@ktu.edu

As the amount of data increases each year, these amounts of data become increasingly difficult to analyze. Currently, a variety of different machine learning algorithms are proposed for data analysis to help make different versions, and research and other activities require solutions. Probably the two most significant types of machine learning are supervised learning and unsupervised learning. If there is no prior knowledge of the data class, unsupervised learning is required. One of the most commonly used forms of unsupervised learning is clustering. Clustering is often described as a particular process that seeks to find data contained in hidden relationships. It is unnecessary to know the class in advance to find these connections, which allows the data in the main groups to be distinguished. Data clustering can be performed using various methods, but they are all divided into four main groups: partitioning methods, hierarchical methods, density-based methods, and grid-based methods. Partitioning methods are described as methods that are flexible and are based on the iterative division of data points into clusters and the subsequent redistribution of these points between groups. The most commonly known and one of the most widely used methods is k-means. Hierarchical clustering is a recursive partitioning of a dataset into successively smaller clusters. Hierarchical methods work by creating a hierarchy of groups. Density-based clustering is a nonparametric approach where the clusters are high-density areas of the density p (x). Grid-based clustering is the last class of clustering methods. This class of methods works by dividing the entire data space into a grid structure with a certain number of cells. Clustering is then performed with these cells instead of individual points, and for this reason, It can significantly reduce the computation time. This work aims to compare

different groups of clustering methods and particular methods using different data sets and evaluate their performance. This work also seeks to include methods that are better known to everyone and much less commonly used. Finally, this work will help to provide some guidance on when specific methods are best suited.

Decomposition Problems in the Development of Complex Information Systems

Audronė Lupeikienė

Institute of Data Science and Digital Technologies
Vilnius University
audrone.lupeikiene@mif.vu.lt

System decomposition, in other words, architecture selection, is an essential decision that determines the success of the system created. The problem is critical as modern information systems expand in scope and complexity. This means that decomposition must be justified and quantitatively evaluated. However, this is not the usual rule in practice, especially in the information systems early architecting stage. Moreover, only very few research papers on the subject have been published.

This is the position paper and summarizes the results of research on the decomposition, covering not only the area of information systems, but area of systems as well. It highlights the problems which should be addressed and resolved by the developers of complex information systems. Some from these problems were identified as a result of the study of scientific literature, others – through an in-depth analysis of the gained experience. The paper contributes to information systems decomposition theory by proposing a holistic approach to decomposition problems.

COVID-19 Infection in Lithuania: Analysis of Social-Economic Consequences

Jurgita Markevičiūtė[1], Jolita Bernatavičienė[2], Rūta Levulienė[1], Viktor Medvedev[2], Povilas Treigys[2]

[1] Institute of Applied Mathematics
Vilnius University
[2] Institute of Data Science and Digital Technologies
Vilnius University
jurgita.markeviciute@mif.vu.lt

Using statistics, econometrics, machine learning, and functional data analysis methods, we evaluate the consequences of the lockdown during the COVID-19 pandemics for wage inequality and unemployment. We deduce that these two indicators mostly reacted to the first lockdown from March till June 2020. Also, analysing wage inequality, we conduct analysis separately for males and females and different age groups. We noticed that young females were affected mostly by the lockdown. Nevertheless, all the groups reacted to the lockdown at some level.

Deep Learning in Alzheimer's Disease

Rytis Maskeliunas

Faculty of Informatics
Kaunas University of Technology
rytis.maskeliunas@ktu.lt

Deep learning has shown tremendous potential in medical applications, not excluding hard to detect symptoms of Alzheimer's and related diseases. Accessibility of data deriving from neuroimaging techniques, such as structural and functional MRI, positron emission tomography and imaging genetics allowed a breakthrough in clinical decision support. The presentation showcases a range of models and applications, discussing challenges and implications within this topic.

User Behavior Based Host-Level Intrusion Detection Using Deep Neural Network

Dalius Mažeika, Elonas Ševiakovas

Department of Information Systems
Vilnius Gediminas Technical University
dalius.mazeika@vilniustech.lt

Intrusion detection is a relevant field of information security, and different artificial intelligence methods are used to identify cyberattacks and anomalies in the networks and hosts. In this research, we address the problem of identifying host-level intrusion detection through time-series data analysis of user behavior. Data such as TCP/IP connections, size of transferred data, and running processes in the host were analyzed. A specialized tool was developed to build a dataset from Windows-based desktop by gathering data of Windows users' normal and abnormal behaviors. The following unauthorized actions as permission escalation, transferring of sensitive user data, SSH service launching, or session opening were treated as intruder activities. Gathered data was proceeded using MD5 feature hashing and normalized, applying min-max scaling or L2 norm depending on the data type. A deep learning approach using LSTM autoencoder was implemented for host intrusion detection. The model was trained until 100 epochs using a dataset collected during two days, while the third day's data were used for model testing. Analysis of the resulting accuracy of the model was performed, and the highest accuracy of 78.57% was achieved when nine records grouped the data. Finally, results were compared with the public dataset ADFA-LD, and corresponding conclusions were made.

Developing an ANFIS-Based Model to Predict Web Services QoS/QoE

Jolanta Miliauskaitė[1], Diana Kalibatiene[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Vilnius Gediminas Technical University
diana.kalibatiene@vilniustech.lt

Nowadays, various stakeholders have provided plenty of web services (WS) with similar functionality to fulfill various organisations' increasingly sophisticated business needs. As a solution, some authors propose using the non-functional properties of Quality of Service (QoS) and user's needs of Quality of Experience (QoE) in planning WSs. The most common QoS description is based on non-functional attributes (like response time, throughput, etc.) on the technological level. They are objective features that are not explicitly related to the user's interests but depend on their implementation and their development quality. QoE attributes (like, cost, reputation, etc.) are subjective characteristics obtained from the user's subjective evaluation of an item at a business level and depend on the morale, interests, and other subjective factors. Consequently, all mentioned attributes, i.e., objective and subjective, should be considered when predicting WS quality. However, while QoS attributes are expressed in a numerical form and can be used to determine QoS using data-driven prediction methods, QoE attributes can be expressed in a non-numerical/linguistic form. Different users have diverse requirements for various QoSs. Consequently, we need a hybrid fuzzy-based reasoning approach and system, which can deal with numerical and linguistic data inputs and enable us to express user experience through vague concepts. To address these issues, we propose a new fuzzy-based reasoning approach for predicting WS quality with the adaptive neuro-fuzzy inference system (ANFIS) suitable for processing numerical and linguistic data input. The proposed approach was implemented as a prototype, and two experiments were conducted. The results of the experiments comparison show a good performance and efficiency of the proposed approach for predicting WS quality. Consequently, it can be considered a suitable tool for predicting WS quality.

World of Code: Enabling a Research Workflow for Mining and Analyzing the Universe of Open Source VCS Data

Audris Mockus

Department of Electrical Engineering and Computer Science
University of Tennessee, USA
audris@utk.edu

Open source software (OSS) is essential for modern society and, while substantial research has been done on individual (typically central) projects, only a limited understanding of the periphery of the entire OSS ecosystem exists. For example, how are the tens of millions of projects in the periphery interconnected through technical dependencies, code sharing, or knowledge flow? To answer such questions, we: a) create a very large and frequently updated collection of version control data in the entire FLOSS ecosystems named World of Code (WoC), that can completely cross-reference authors, projects, commits, blobs, dependencies, and history of the FLOSS ecosystems and b) provide capabilities to efficiently correct, augment, query, and analyze that data. Our current WoC implementation is capable of being updated on a monthly basis and contains over 18B Git objects. To evaluate its research potential and to create vignettes for its usage, we employ WoC in conducting several research tasks. In particular, we find that it is capable of supporting trend evaluation, ecosystem measurement, and the determination of package usage. We expect WoC to spur investigation into global properties of OSS development leading to increased resiliency of the entire OSS ecosystem.

Somatic and Mental Symptoms in Patients with Chest Pain: A Cross-Sectional Study

Kristina Morkūnaitė[1], Vytenis Tamakauskas[2], Rytis Leonavičius[1], Darijus Skaudickas[1], Vincentas Veikutis[2]

[1] Lithuanian University of Health Sciences
[2] Institute of Cardiology, Lithuanian University of Health Sciences
vincentas.veikutis@lsmuni.lt

The aim of this study was to review the relationship between panic disorder and CAD in patients with chest pain was to identify characteristics of the chest pain associated with the presence of panic disorder, to determine the strength of the association between panic disorder and CAD, and to determine the association between panic disorder and known cardiovascular risk factors.

Measuring the Quality of Synthetic Speech

Gediminas Navickas, Gerda Ana Melnik-Leroy, Povilas Treigys

Institute of Data Science and Digital Technologies
Vilnius University
gediminas.navickas@mif.vu.lt

Measuring the quality of synthetic speech is very important for the development of high-quality Text-To-Speech (TTS) systems and for defining, what quality is enough for particular user groups. Sometimes a small increase in quality may have a great cost in terms of computing resources and training data. Usually the quality is measured using intelligibility and comprehension tests, but they have many limitations and are no longer suitable to test modern TTS systems. Thus, developing and applying more sensitive measures, with well-defined and experimentally controlled conditions, is necessary in order to bring further advances in the field of TTS.

In this study, we raise the question of synthetic speech evaluation, using a controlled experimental paradigm. First, we evaluate experimentally the perceptual differences between objectively different synthesized speech qualities. Second, we compare the perception of two groups of listeners: sighted and blind participants.

The experimental paradigm is based on a modified AX (same-different) discrimination task, in which participants hear two samples of synthesized speech in each trial and they have to answer whether they sound same or different.

The results show that experimental paradigm can be used for the evaluation of synthetic speech quality and also for defining the level of perceptual accuracy for particular user groups.

Empirical Analysis of Selected Blockchain Simulators

Remigijus Paulavičius, Ernestas Filatovas

Institute of Data Science and Digital Technologies
Vilnius University
remigijus.paulavicius@mif.vu.lt

In recent years, various blockchain-based solutions have been created. However, the lack of tools to evaluate blockchain systems may limit the development of the field. Many benefits of blockchains can be demonstrated only at large scales, e.g., using thousands of nodes. Therefore, the investigation of different implementations and design choices is complicated and hardly feasible on real blockchains. Meantime, blockchain simulators give the possibility to reproduce complex real-world systems at a low cost. In this talk, we present the first systematic review and empirical analysis of existing blockchain simulators. Most of these simulators are readily extensible and can be used to test the performance of blockchains with different settings and parameters on a single computer. The features and limitations of selected simulators are summarized and experimentally validated. Finally, possible future research directions in the field are highlighted.

The Analysis of Impact of Noise on Hyperspectral Unmixing Algorithms

Vytautas Paura, Virginijus Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University
vytautas.paura@mif.stud.vu.lt

Over the past decades, many different methods have been proposed to solve the linear or nonlinear mixing problems in hyperspectral unmixing. A few main approaches emerged over the years, such as: Nonnegative Matrix Factorization, Linear mixture modelling and Autoencoder networks. Each type of method implements a different way to find the number of endmembers, their signature spectras and whole abundance matrices in order to fully unmix the hyperspectral data into its parts. These different methods are created using various algorithms that may be influenced more by the amount of noise the hyperspectral images have. Due to the sensors used in hyperspectral cameras the quality of the data is influenced by a lot of different factors like light amount, environmental and atmospheric effects, and electrical noise. All of these factors influence the amount of noise that is gathered by hyperspectral sensors. We performed an experiment in order to analyse the influence, of the amount of noise in hyperspectral data, has on the algorithm's abilities to unmix the hyperspectral data. An array of algorithms are tested using the created experiment to determine the best performing algorithms. Improvement directions on the best performing algorithm are presented to improve the results independently of the amounts of noise in the images.

Electric Vehicle Energy Consumption Modelling and Estimation

Linas Petkevičius[1], Simonas Šaltenis[1],
Alminas Čivilis[1], Kristian Torp[2]

[1] Vilnius University
[2] Aalborg University, Denmark
linas.petkevicius@mif.vu.lt

The growth of electric vehicles raises new challenges. For example, long-distance route planning involves not only computing a suggested route, but also planning the charging stops along the route. This in turn relies on the estimation of energy consumption on the edges of the road network – the main focus of our work. Under the assumption that energy use is predicted from the estimated speed profile and other contextual characteristics, such as weather information and slope, the prognostic model is constructed. In particular, we investigate deep-learning models that are built from EV tracking data. We present different experimental setups and investigate the accuracy and other characteristics of the proposed ML models for EV energy-consumption estimation. Further, we demonstrate model behavior within the interpolated and extrapolated conditions.

Knowledge-Base Enriched Word Embeddings for Social Media Text Analysis

Mindaugas Petkevičius, Daiva Vitkutė-Adžgauskienė

Vytautas Magnus University
mindaugas.petkevicius@vdu.lt

Social media text analysis is important in different application areas – for example, web resource mining for marketing needs, fact and event detection for electronic space security tasks, etc. Deep learning techniques are used widely for text analysis today, and the quality of such analysis is greatly determined by the quality of word embedding models. Typically, word embeddings rely on word context, however for brief social texts (user comments, forum posts, etc.) this context is often very limited, and texts used for learning the embedding model are frequently incomplete and ambiguous. On the other side, using word embeddings generated from regular and public domain texts, such as Wikipedia, does not reflect the specifities of social space language, where the dictionary contains a high percentage of irregular and misspelled words, as well as some jargon. Therefore, when creating word embeddings for social text analysis tasks, we suggest leveraging extra knowledge in order to understand text correctly and to recognize needed patterns in a precise way. For this purpose, in addition to employing social media texts for developing word embedding models, data from such knowledge bases can be used: 1) Wordnet – lexical ontology for pulling semantically related words (synonyms, hypernyms, hyponyms, etc.) closer to each other; 2) Domain-specific semantic data sources, for example classified NER knowledge bases; 3) Dictionaries related to social media language specifics. In this study, we demonstrate the advantage of enriched embeddings for solving a simple intrinsic language task of word similarity evaluation, using FastText embeddings, enriched and modified with additional semantic information, and a complex extrinsic task of event detection in social texts, using a convolutional neural network (CNN) with additional semantic layer. By incorporating an additional semantic layer into the deep

learning model, we can compensate for sparse contextual information found in brief, irregular and misspelled words, and thus achieve more accurate event identification. In our experiments we used 300-dimensional FastText word vectors, that demonstrated competitive performance in a word semantic-relatedness tasks, and LitWordNet ontology for additional semantic information. Retrofitting approach was used for adding semantic LitWordNet information to pre-trained word vectors by pulling semantically related words closer together. Furthermore, the embedding model was enriched by adding words from specialized social media language dictionaries. When comparing Pearson correlation results for traditional embeddings and knowledge-base enriched word embeddings, we observe an increase in Pearson correlation accuracy as well as an increase in F-score for event detection task. The results, of our study show, that enriching word embeddings with additional semantic information from external knowledge bases can be beneficial for both simple intrinsic language tasks, such as word-level similarity evaluation, as well as for complex extrinsic text analysis tasks, such as event detection in social media texts.

Evaluation Metrics for Synthetic Social Media-Derived Texts

Ignas Rudaitis[1], Justina Mandravickaitė[2], Danguolė Kalinauskaitė[2], Veronika Gvozdovaitė[1], Tomas Krilavičius[2]

[1] CARD – Centre for Applied Research and Development
[2] Vytautas Magnus University
ignas.rudaitis@card-ai.eu

Recently, machine learning has been employed to solve various tasks related to social networks. As a result, it can be observed a growing need for snapshots of virtual social networks, containing the interaction graphs, the related textual content, or both. However, it has become difficult to meet this need due to data protection regulations. In turn, this issue has sparked interest in synthetic datasets of social media-style texts and graph structures. In our research, we focus on textual content. We analyze different language models (LMs) for synthetic social media data generation. Perplexity is a metric used for LMs evaluation. However, evaluating the output of synthetic social media texts generation is a non-trivial task: on the one hand, it is expected for the LMs to replicate the grammar, the lexicon, and the social dynamics of social media platforms; on the other hand, it is not necessary for the models to accurately reproduce any specific content. Standard metrics of machine translation or natural language generation, such as BLEU or ROUGE, cannot be used for this task since they rely on a naturally supervised process of the task. We propose to use supervised metrics by admixing natural language datasets with phrases from an artificially generated language. This is expected to naturally conform to the tendency for code-switching that is common in social media platforms. In this context, several criteria for LMs evaluation should be considered:

1. Recalling of synonyms for the admixed artificial words and phrases.
2. Capturing conventions when artificial phrases are preferred to the original ones.

3. Preserving of relations within comment threads where artificial phrases are used.
4. Preserving of grammatical logic within artificial phrases.

Based on these criteria, we evaluate several models, including Markov chains of various orders, a character-level LSTM, and a fine-tuned snapshot of GPT-2. We summarize the results based on their correlation with conventional measures, such as perplexity.

Mortality Rate Estimation Models of Patients with Prostate Cancer Diagnosis

Tomas Ruzgas[1], Vytautas Kraujalis[1], Daimantas Milonas[2]

[1] Kaunas University of Technology
[2] Lithuanian University of Health Sciences
tomas.ruzgas@ktu.lt

Prostate cancer is one of the most frequent type of male cancer all around the world, including Lithuania. Prostate cancer diagnosis takes up to 25% of all Lithuanian men cancer diagnosis in the country. Every year, ~3000 new prostate cancer cases are diagnosed and ~500 Lithuanian men die from this illness. It is necessary for a medical professional to be able to distinguish a fatal and non-fatal cancer in time, statistical methods and models could be implemented to help with this case and in our work we will try to implement those methods and models using data from "Kauno Klinikos" clinic (Kaunas, Lithuania). During the research we used well known Kaplan-Meier survival curves as well as compared 2 best known hazard estimation models: Cox and Fine-Gray models. In this dataset, there were 56 deaths reported from cancer specific causes and 294 from other causes, the median age of a patient was 64 years (n - 2410). During the analysis of Kaplan-Meier survival curves, we found the worst survival prognosis associated with patients, who's lymph nodes were damaged by cancer or patients with 5 metastatic lymph nodes. It was also discovered that patients with 1 or 2 metastatic lymph nodes are much more likely to experience death from one of the causes – cancer specific cause or other causes than men with 4 metastatic lymph nodes. Significant hazard ratio was also found between men, who's cancer is developing in the prostate area and men, who's cancer has already spread outside the prostate cancer, with the later one having the worse survival prognosis. Patients with cancer damaged lymph nodes have a higher mortality rate than men with untreated lymph nodes only from cancer specific causes while the hazard ratio linked with man's age was found significant only in deaths from other causes. Fine and Gray hazard estimation model distinguished less significant risk factors and usually the

hazard ratios were reported smaller than the ones in the Cox model. Training and testing datasets were used to test the performance of both models and Cox model was found to be optimal on both datasets in response to the ROC curves analysis.

Fuzzy and CMMN Based Dynamic Software Project Management Process Modelling and Simulation

Šarūnė Sielskaitė, Diana Kalibatienė

Vilnius Gediminas Technical University
sarune.sielskaite@vilniustech.lt

The Project Management Institution defined project management as the use of specific knowledge, skills, tools, and techniques to deliver value to people. However, the needs of those people are growing, so naturally, software projects are getting more complex and challenging to implement. Consequently, practitioners and academicians propose different approaches to improve the successful implementation of software projects. One of the project success factors is the appropriately chosen software project management methodology and its adaption to the project type, company, and employees. However, the issue of selecting software project management methodology and its adoption remains relevant to this day. This paper proposes a new approach for choosing and adopting a software project management methodology to a particular context (i.e., the project type, company, employee, etc.). In this approach, we view software project management as a dynamic business process, a complex knowledge-based process embedded in a performance process (e.g., developing software, product development, and so on) that could be modelled and simulated to analyse its changes. Consequently, for this dynamic modelling and simulation, it is proposed using Case Management Model and Notation (CMMN), which is suitable for adaptive case management and decision making by suggesting and keeping people in the manager position. The case model focuses on real and rapidly changing information and relationships relevant to the constantly evolving software project management process. One of the main advantages of the CMMN is that it provides an opportunity to illustrate discrete events, thus allowing case handlers to decide for themselves whether a task is relevant to the execution process and whether to execute it in the software project management process. Moreover, since software pro-

ject management is fuzzy by its nature (i.e., resources, finance budget, course of tasks change), a fuzzy inference system seems to be helpful to model those uncertainties in the form of rules and predict the flow of software project management process. The proposed Fuzzy and CMMN based dynamic software project management process modelling and simulation approach was implemented as a prototype and evaluated with an industry case study. The results showed that the proposed approach is implementable and can be used for software project management simulation.

Investigation of Wireless Sensor Networks Protocols Performance

Julius Skirelis, Dalius Navakauskas

Vilnius Gediminas Technical University
julius.skirelis@vilniustech.lt

Wireless Sensor Networks (WSN) being a part of Internet of Things (IoT) technology is often used as data sensing and data acquiring layer. In the presented performance analysis, the dynamically reconfigurable Mobile Ad-hoc Network (MANET) with parameters of a real system is simulated and investigated. The main feature of MANET leading to its selection for investigation – it does not have any centralized management mechanism, nor strict topological infrastructure, hence each node acts as a router for remaining nodes in the network. Often WSN nodes transfer the data in open field, therefore adverse weather conditions drastically impact MANET performance, in a certain environments different weather conditions may cause delays or even complete data loss. To overcome this issue, variety of different transmission protocols are developed, each of them is based on different mechanisms, thus possesses distinct advantages and disadvantages.

The communication protocol between the sensors and the gateway is the main object of analysis. Four different tree and mesh topology-based protocols were chosen for investigation: 1-Hop, ODMRP, HLMRP and DSR. Simulation and analysis is performed using MATLAB software package and MANET toolbox. Real RF sensors parameters were used for the simulation. Three different RF disturbance models were substituted: fog, rain and snow. Additionally, the baseline performance of each protocol was investigated without any disturbance applied.

Total of 16 different time limited simulations were performed and average absolute number of dropped data packets were analyzed. Finally, the analysis of the overall protocol performance over all disruptions as the ratio of the dropped packets to total sent packets was performed.

The performance analysis of four different WSN protocols simulation results confirm that:

- in the Radio Frequency WSN for IoT use case the natural packet drops are inevitable;
- HLMRP data communication protocol performs the best of all considered protocols.

Deep Learning Models for Hate Speech Detection

Milita Songailaitė, Eglė Kankevičiūtė, Justina Mandravickaitė, Danguolė Kalinauskaitė, Tomas Krilavičius

Vytautas Magnus University
danguole.kalinauskaite@vdu.lt

We discuss an experiment on comparison of deep learning models for hate speech detection. Online hate speech is assumed to be an important factor in political and ethnic violence. Therefore, media platforms are pressured to timely detection and elimination of hate speech. This tendency led to increasing efforts in terms of hate speech detection, and several hate speech detection models have been developed. As hate speech is not only a complex phenomenon that is difficult to detect but even its definitions vary in different studies, the purpose of this experiment is to compare selected hate speech detection models for English from the perspective of inter-annotator agreement. It is widely used in corpus linguistics, computational linguistics, discourse studies, etc. to evaluate the reliability of an annotation process. In our experiment case, inter-annotator agreement metrics have been used to evaluate how the selected models "agree" in terms of annotation of hate speech instances. For model comparison, we used an English dataset from HASOC 2019 shared task. For comparison, we selected the following models for hate speech detection: CNN+GRU, HateBERT, BiRNN, BiRNN-Attn, BiRNN-Scrat, BERT-HateXplain. For assessing inter-annotator agreement Fleiss' kappa and Krippendorff's alpha metrics were used as they allow multiple comparisons.

Music Similarity Evaluation Using Machine Learning Methods and Audio Signal Profiles

Milita Songailaitė, Tomas Krilavičius

Vytautas Magnus University
milita.songailaite@stud.vdu.lt

Nowadays the music is more accessible to us than ever before. With the increased popularity of online music streaming companies, people find themselves spending more and more time while choosing the songs they actually like. This poses a problem of a fast and accurate music recommendation method, which would let the user ignore the large quantities of songs and choose precisely what he likes. In this work, we present a method to compare music based entirely on its audio signal properties. For this, we used three different approaches (Gaussian Mixture Models, Dynamic Time Wrapping and Autoencoders) to calculate the similarity between the given signals. All three experiments were performed on a database consisting of 2511 most popular songs from 10 different genres. The methods were evaluated by comparing algorithms results with the music similarity results given by the experts. After the evaluation, it turned out that the most accurate method was the Gaussian Mixture models. The rest of the models gave similarly worse similarity scores.

Cancerous Tissue Detection Using Dynamic Contrast-Enhanced MRI Data for Prostate Region

Roman Surkant[1], Justinas Jucevičius[1], Povilas Treigys[1], Jolita Bernatavičienė[1], Mantas Trakymas[2], Ieva Naruševičiūtė[2]

[1] Institute of Data Science and Digital Technologies Vilnius University
[2] National Cancer Institute
roman.surkant@mif.stud.vu.lt

Prostate cancer is one of the leading causes of cancer death worldwide. Among males, prostate cancer has the second highest incidence rate after lung cancer. Although death rates have been decreasing in some countries, it remains a considerable disease affecting many patients and early diagnosis and treatment are critical. Preliminary identification of cancer involves biopsy PSA protein screening, elevated levels of which indicate an increased likelihood of prostate cancer. Unfortunately, such testing is invasive and prone to false-negative and false-positive results, so a less invasive and more reliable procedure is needed. Currently, evaluation is done using different types of imaging, each having own acquisition methods and purpose, and the final diagnosis is formulated based on all of them in conjunction. Dynamic contrast-enhanced (DCE) images, one of such imaging types, is unique in a way that it shows the flow of a contrast medium injected into patient's bloodstream over time highlighting tissues that have higher vascular density, for instance, cancerous growth. This work is dedicated to investigating the usage of DCE imaging to detect malignant tissues by constructing time-signal intensity curves showing stepwise changes in enhancement over time in different prostate regions.

Several Stochastic Models for Non-Life Insurance Business

Jonas Šiaulys

Institute of Mathematics
Vilnius University
jonas.siaulys@mif.vu.lt

The so-called risk renewal model with certain supplements describing the investment environment is commonly used to describe non-life insurance business. The main part of such models is described by the flow of claims. If claims are considered to be identically distributed random variables, then we get so-called homogeneous risk renewal model. If we suppose that the claims are not necessarily identically distributed, then we obtain the inhomogeneous risk renewal model. The report will review several results on the critical characteristics of the inhomogeneous risk renewal model.

On the Computed Tomography Image Data to Diagnose Pancreatic Cancer Using Machine Learning

Aušra Šubonienė[1], Olga Kurasova[1], Viktor Medvedev[1], Aistė Kielaitė-Gulla[2], Artūras Samuilis[3], Džiugas Jagminas[4], Kęstutis Strupas[2], Gintautas Dzemyda[1]

[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, Department of Radiology, Nuclear Medicine and Medical Physics, Faculty of Medicine
Vilnius University
[4] Faculty of Medicine
Vilnius University
ausra.suboniene@mif.vu.lt

Medical imaging data, which is suitable for solving segmentation problems, are difficult to obtain due to data sensitivity issues and the effort that is required to make ground truth segmentations. In order to increase the robustness of results by including more medical images, multiple datasets are often combined. However, challenges arise when trying to combine such datasets from different sources. Populations of patients that differ by age and other conditions could affect the results. Also, there might be different approaches to segmentation and its accuracy. Experts can segment medical images as true to anatomical structures as possible, or they might include some surrounding tissues in order to speed up manual segmentation. Also, rough region boundaries can be used instead of segmentations. Lastly, there can be different diagnostic devices used, which might result in different pre-processing of images.

Data sources. Due to data sensitivity and effort that is needed to anonymise images, there is a lack of publicly available pancreatic cancer data. Most publicly available medical data is without segmentation by experts, and images that do have some anatomical structures segmented are scarce. Currently, the largest public collections of computer tomography (CT) images of pancreatic cancer are available are the Can-

cer Imaging Archive (TCIA) dataset and Medical Segmentation Decathlon dataset. The Medical Segmentation Decathlon dataset consists of 421 portal-venous phase 3D CT scans. Segmentations of both the pancreatic parenchyma and pancreatic mass (cyst or tumour) are provided, although done as ROI only, which makes the segmentation process quicker but less accurate. TCIA dataset consists of 82 abdominal contrast enhanced 3D CT scans with slice thickness between 1.5−2.5 mm. Manual segmentations were done only to segment the pancreas. Here we also analyse the dataset which was acquired in the Vilnius University Hospital Santaros Klinikos. Segmentations that were provided by the experts consist of healthy pancreas, pancreatic cancer and pancreatic duct, which can be confused with pancreatic cancer by machine learning algorithms due to similar intensity of pixels. When combining publically available datasets segmentation differences need to be taken into account and some additional manual segmentation might be needed.

Differences in populations of patients. Due to the limited availability of medical images and pancreatic cancer being more common later in life, it is difficult to achieve equal coverage of all age categories of pancreatic cancer patients. This can be partially solved by increasing the dataset size of categories by retrospective analysis of computer tomography images provided by the medical institution, that have been filtered by the desired conditions.

Issues in the quality of the segmentation. Manual segmentations of computer tomography images can be prepared using different approaches to segmentation and its accuracy. Experts can segment medical images as true to anatomical structures as possible, or they might include some surrounding tissues in order to speed up manual segmentation. Also, rough region boundaries can be used instead of segmentations. Data pre-processing might be done by removing fat tissue based on the values of Hounsfield units. This results in holes and irregular edges of segmented regions. This creates additional problems when trying to unify segmentations across multiple datasets and might reduce the accuracy of segmentations when using machine learning.

Different diagnostic devices. Lastly, there can be different diagnostic devices used, which might result in different pre-processing of images. Even after pre-processing, these different diagnostic devices can have unique artefacts in computer tomography images that can reduce segmentation accuracy when combining multiple datasets.

P2Y12 Inhibitors (Ticagrelor) Adverse Effect Prediction Depending on Clinical Data and Genetic Factors in Patients after Acute Coronary Syndrome

Vytenis Tamakauskas[1], Vacis Tatarūnas[1],
Nora Kupstytė-Krištaponė[2], Vincentas Veikutis[1]

[1] Institute of Cardiology, Lithuanian University of Health Sciences
[2] Lithuanian University of Health Sciences
v.tamakauskas@gmail.com

This study aimed analyze effectiveness, advert effects and genetical factors of P2Y12 inhibitors (Ticagrelor) in patients after acute coronary syndrome. Major goal is to determine factors that will predict effectiveness of P2Y12 inhibitors and prevent most common side effects such as dyspnea, bleeding or coronary stent thrombosis.

Scalable Trust Region Bayesian Optimization with Product of Experts

Saulius Tautvaišas, Julius Žilinskas

Institute of Data Science and Digital Technologies
Vilnius University
saulius.tautvaisas@mif.vu.lt

The global optimization of high-dimensional black-box functions is an important component of modern machine learning. While Bayesian optimization (BO) has become an effective approach for low-dimensional problems, it scales poorly to high-dimensional problems with large number of observations. Gaussian Process (GP) based BO is typically limited to only a few thousands of observations, because of its cubic training cost. In this paper we propose a generalized Product of Experts based Trust Region Bayesian optimization (gPoETRBO) algorithm to address the scalability issues of the standard GP model. In our experiments, we show that our proposed algorithm achieves similar performance to other state-of-the-art algorithms but is more scalable and computationally more efficient.

Generating of Non-Regular Instances of Semidefinite Programming Problems

Tatiana Tchemisova

Mathematical Department
University of Aveiro, Portugal
tatiana@ua.pt

Semidefinite programming (SDP) deals with the problem of minimising linear functions subject to linear matrix inequalities (LMIs) and belongs to conic optimisation. A wide variety of nonlinear convex optimisation problems can be formulated as problems involving LMIs, and hence efficiently solved using recently developed interior-point methods. Semidefinite programming has been recognised in combinatorial optimisation as a valuable technique for obtaining bounds on the solution of NP-hard problems. It provides important numerical tools for analysis and synthesis in systems and control theory, robust optimisation, computational biology, systems and control theory, sensor network location, and data analysis, among others. Regularity is an important property of optimisation problems. Various notions of regularity are known from the literature, being defined for different classes of problems. Usually, optimisation methods are based on the optimality conditions, that in turn, often suppose that the problem is regular. The absence of regularity leads to theoretical and numerical difficulties, and solvers may fail to provide a trustworthy result. Therefore, it is very important to verify if a given problem is regular in terms of certain regularity conditions and in the case of nonregularity, to apply specific methods. On the other hand, in order to test new stopping criteria and the computational behavior of new methods, it is important to have an access to sets of reasonably-sized nonregular test problems. We present a generator that constructs nonregular SDP instances with prescribed irregularity degrees and a database of nonregular test problems created using this generator. Numerical experiments using popular SDP solvers on the problems of this database permit us to conclude that the most efficient solvers are not efficient when applied to nonregular problems.

Modeling of Crystallization Conditions for Organic Synthesis Product Purification Using Deep Learning

Mantas Vaškevičius, Jurgita Kapočiūtė-Dzikienė

Informatics Faculty, Vytautas Magnus University
mantas.vaskevicius@vdu.lt

Crystallization is an important purification technique of solid products in a chemical laboratory. One of the drawbacks is that a correct solvent must be selected for the procedure. To speed up solvent search we propose a deep learning method to directly predict solvent labels. An open-source collection of chemical reactions, comprised of 3.7 million raw reactions and synthesis procedures, has been used to create the dataset to research machine learning algorithms. We have made use of structured synthesis procedures to extract the solvent names used in the crystallization step of the syntheses. We have tested two vectorization methods with different parameters – extended-connectivity fingerprints (ECFP) and ECFP autoencoders. Feed-forward and LSTM neural networks have been used to train multi-label classifiers to predict the solvent used in the crystallization step of the synthesis procedure. The most accurate optimized model can predict solvent labels with an accuracy of 0.87 ± 0.004. We provide analysis of vectorization and deep learning methods of a novel method for prediction of solvent used in the crystallization step of organic synthesis. Research results may be used to accelerate R&D processes in the laboratories. Additionally, a method for modeling of reaction mixture crystallization conditions invariant to the reaction type has not been previously demonstrated.

Impact of Images Quality Variety and Resizing Level on Eye Fundus Optic Disc Segmentation

Sandra Virbukaitė, Jolita Bernatavičienė

Institute of Data Science and Digital Technologies
Vilnius University
sandra.virbukaite@mif.vu.lt

Various eye diseases such as glaucoma, diabetic retinopathy and hypertension can be diagnosed using eye fundus images. Therefore, image analysis is necessary. Here, the different parts of an eye such as blood vessels, macula, optic disc, and optic cup may be the objects of interest depending on the disease. In eye fundus images analysis, image segmentation is one of the main steps. At this stage, different objects in the image are distinguished and defined, and thus assigned to different object classes. With the rapid development of convolutional neural networks in image processing, deep learning methods have achieved great results in automated image segmentation. Applying various deep learning algorithms, the image quality plays an important role. In this research we analyzed a few different publicly available datasets. Each dataset consists of different quality images as these have been captured by different non-stationary digital eye fundus cameras. The images vary in resolution, brightness, and visualization. Several images pre-processing scenarios have been applied to evaluate an impact of these images' quality on image segmentation. An impact of images resizing level has been evaluated as well. For these evaluations we applied the most popular medical images segmentation autoencoder named U-Net. Optic disc has been chosen as an object of interest.

Telecommunication Customer Churn Prediction Using Machine Learning Methods

Monika Zdanavičiūtė, Rūta Juozaitienė, Tomas Krilavičius

Vytautas Magnus University
monika.zdanaviciute@card-ai.eu

These days telecommunication sector has grown significantly due to the use of smart technologies, and it is likely to continue to grow. The main resource of telecommunications companies is customers, but due to the relatively high level of competition in this field, most customers are not tied to a single service company. To understand the key factors contributing to customer churn rate, we have analysed the real data of telecommunication company. The half-year data consisted of information on 23 559 users, 14 561 payments and 333 651 calls. The main contribution of our work was to develop a churn prediction model which identifies customers who are most likely subject to churn. We performed experiments using k-Nearest Neighbours, Support Vector Machine, Decision Trees, Random Forest, Naïve Bayes classifiers and the Cox proportional hazard model with time-varying covariates. Results show that the Cox regression model with time-varying covariates was superior to classical classification methods because it can take into account static user parameters and reflect their changes over time.

For notes

12th Conference on

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

Partner

General sponsors

VTeX

NOVIAN

Main sponsor

asseco

Sponsor

BALTIC AMADEUS