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

10th International Workshop on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

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

November 29 – December 1, 2018

VILNIUS UNIVERSITY PRESS
Vilnius, 2018

Co-Chairmen:

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

Programme Committee:

Prof. Juris Borzov (Latvia)
Prof. Albertas Čaplinskas (Lithuania)
Prof. Robertas Damaševičius (Lithuania)
Prof. Janis Grundspenkis (Latvia)
Prof. Hele-Maja Haav (Estonia)
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
Prof. Olga Kurasova
Dr. Viktor Medvedev
Dr. Jolanta Miliauskaitė
Laima Paliulionienė
Dr. Martynas Sabaliauskas

Contacts:

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

https://doi.org/10.15388/DAMSS.2018.1

ISBN 978-609-07-0043-3

Preface

DAMSS-2018 is the jubilee 10th international workshop on data analysis methods for software systems, organized in Druskininkai, Lithuania, at the end of the year. The same place and the same time every year.

Ten years passed from the first workshop. History of the workshop starts from 2009 with 16 presentations. The idea of such workshop came up at the Institute of Mathematics and Informatics. 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 81. The number of registered participants is 113 from 13 countries.

In 2010, the Institute of Mathematics and Informatics became a member of Vilnius University, the largest university of Lithuania. In 2017, the institute changes its name into the Institute of Data Science and Digital Technologies. This name reflects recent activities of the institute. The renewed institute has eight research groups: Cognitive Computing, Image and Signal Analysis, Cyber-Social Systems Engineering, Statistics and Probability, Global Optimization, Intelligent Technologies, Education Systems, Blockchain Technologies.

The main goal of the workshop is to introduce the research undertaken at Lithuanian and foreign universities in the fields of data science and software engineering. Annual organization of the workshop allows the fast interchanging of new ideas among the research community.

Even 11 companies supported the workshop this year. This means that the topics of the workshop are actual for business, too. Topics of the workshop cover big data, bioinformatics, data science, blockchain technologies, deep learning, digital technologies, high-performance computing, visualization methods for multidimensional data, machine learning, medical informatics, ontological engineering, optimization in data science, business rules, and software engineering. Seeking to facilitate relations between science and business, a special session and panel discussion is organized this year about topical business problems that may be solved together with the research community.

This book gives an overview of all presentations of DAMSS-2018.

Supported by:

General sponsors

Algoritmų sistemos
www.algoritmusistemos.lt

Neurotechnology
www.neurotechnology.com

Western Union Processing Lithuania
www.westernunion.com

Main sponsors

Asseco Lithuania
asseco.lt

BAIP
www.baip.lt

VTeX
vtex.lt

Sponsors

Baltic Amadeus
www.baltic-amadeus.lt

Monet LT
www.monet.lt

NRD CS
www.nrdcs.lt

CoinGate
coingate.com

Visoriai Information Technology Park (VITP)
www.vitp.lt

Real-Time Identification of Ventricular Tachycardia Incident Using Multi-Modal Classifier

V. Abromavičius, A. Serackis

Department of Electronic Systems
Vilnius Gediminas Technical University
arturas.serackis@vgtu.lt

The presented investigation faces the problem of Ventricular Tachycardia detection in real-time using single-channel ECG signal continuous automated monitoring technique. Cardiac arrhythmia detection in heart activity related signals, like ECG, has been a center of attention for many researchers. This particular investigation aimed to develop a technique to improve the accuracy of commonly used classifiers in Ventricular Tachycardia incident detection. A technique, proposed in this paper uses a set of six features, estimated from single ECG signal in a real-time manner. Signal features were estimated taking fixed duration signal frame, centered at QRS complex, detected using the Pan-Tompkins detector. Estimated features were used to classify individual heart beats of the signal with five different techniques – Bagging predictors, Random Forest, Committee of Random Forests, and two Adaptive Boosting techniques for J48 and partial decision trees. The contribution of this investigation is a multi-modal classifier, which is based on the fusion of five techniques. The final Ventricular Tachycardia alarm result is determined by rules-based decision and voting of the five classifiers. As an input to the classifiers, we have used the following features: moving average, standard deviation, variance, covariance, kurtosis and additionally proposed a feature, which compares the duration between SQ and QR point in neighboring QRS complexes.

Due to the unusual ECG signal shapes, high signal-to-noise ratio and other artifacts no individual classifier showed exceptional results. However, the experimental investigation showed that the best performance could be achieved by fusing the results of individual classifiers via voting and rule-based decisions.

The performance of the proposed solution was compared on the 2015 PhysioNet/CinC Challenge sample entries data set. The received performance score of the proposed algorithm was comparable to Top 5 performance scores, received by most successful Challenge participants.

Multi-Platform Blockchain – Integration Between Blockchain's

A. Adamonis, E. Filatovas, R. Paulavičius

Institute of Data Science and Digital Technologies
Vilnius University
andrius.adamonis.2@mif.vu.lt

The multi-platform blockchain is a blockchain network integration type. Multi-platform blockchain enables integration between two or more blockchain networks, i.e., information migration between the networks, information audit and other. Confidential data/information protection issue is very relevant these days. One type of data should be publicly available, while sensitive data must be kept in the private servers. According to the data types mentioned above, we have two kinds of blockchain networks: 1) public blockchain (in which data is held publicly); 2) private blockchain (data held privately). This work presents the concept of blockchain networks integration which could help to solve data privacy and data migration issues between the blockchain (or Distributed Ledger Technology) networks. The concept suggests integration between private blockchain network (the network itself is kept in the internal enterprise servers and holds sensitive/confidential data) and public blockchain network (data is distributed through all the participants of the network, holds non-sensitive information). The main idea of the concept is that sensitive data becomes non-sensitive and moves from private to the public blockchain network, while keeping the relation to the data kept in the private blockchain network. The main issues of multi-platform blockchain network briefly reviewed in this work: a safety of data migration between networks, type of integration software and other.

Application of Specific Business Rules in ERP System for Effective Manufacturing Process of Woven Fabrics

E. Arbataitis, D. Dzemydienė

Institute of Data Science and Digital Technologies
Vilnius University
edga.arba@gmail.com

The aim of manufacturing process effective management requires special methods of knowledge management integrated in multilevel architecture of components of enterprise planning system (ERP). The representation of specifity of manufacturing processes of the textile fabrics requires the integration of specific rules of woven fabrics compatible with the whole productivity optimization process. The studies of textile manufacturing process are trying to solve the optimization problems from theoretical and practical viewpoints, but concrete management process have specifity which can be extracted and applied in concrete enterprise.

Our research work aims to integrate business management rules into the enterprise resource planning and management system architecture based on Service Oriented Architecture (SOA). Service-style architecture can combine applications regardless of the platform (i.e., the operating system or the server architecture used). A programming language is available, which functionality is available through web services. With this study we are trying to fill the gap related with knowledge based methods for effective manufacturing process of woven fabrics. It is discussed not only on application of knowledge base set of rules used for manufacturing process optimization, but also how to construct the framework in which knowledge elements, facts, constraints, and rules of production process is derived using knowledge capturing methodology integrated in production system. We are suggesting to apply this framework for manufacturing process optimization of woven fabrics by solving key issues related to rules and data for product design and process planning. This could help knowledge based system self learn and change in order to cope with ever changing business requirements.

Data Fusion in Big Data Decision Making

J. Arsenjeva, V. Medvedev, G. Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University
jaroslava.arsenjeva@mif.vu.lt

The main idea of data fusion is to merge data from different sensors into one common system. The benefits of using multiple sensors instead of one are undoubtful: for example, if a person is healthy, he has five sensors (eyes, ears, nose, skin and tongue) and can make a better decision than a blind person with 4 sensors (minus the eyes). The same would work for the data mining with artificial sensors: the more information can be obtained the better solutions for a problem any model can propose. Some application ideas for data fusion can be in the medical field (patient monitoring: heart rate, heart pressure and other conditions), military field (target acquiring), agricultural field, etc. The main difficulties encountered while using data fusion: inconsistency of accuracy (when some noise occurs in the data due to environmental changes or imperfection of the sensor itself), conflict of interest, fusion of modalities (various formats of data must be transformed into a common framework), impreciseness of timing, data overload.

Processing data from multiple sensors leads to increasing the amount of information obtained and classifying that information into the Big Data category. Big Data is usually challenging to process and hard to understand for a common user when making a decision. To overcome these difficulties aggregation methods such as data clustering and data fusion can be used making data processing more efficient and accurate. The data fusion model with several levels of focusing selects the necessary inputs, creates a model that would take them into account and evaluates the effectiveness of the output. Nowadays many new possibilities for this model have emerged due to new trends, e.g. Big Data, Internet of Things and Industry 4.0.

Identification of Dynamic Parameters and Velocity Control of a Moving IoT Node Using a Single Ranging Measurement Source

K. Bagdonas, A. Venčkauskas

Department of Computer Science
Faculty of Informatics
Kaunas University of Technology
kazimieras.bagdonas@ktu.lt

A novel method to identify the parameters of the dynamic system is presented in this paper. We consider Internet of Things (IoT) node in motion and extract its velocity and acceleration parameters from continuous ranging measurements performed with a single IoT stationary node via Round Trip Time (RTT) estimation. Obtained parameters are then used as inputs to a simulated Hybrid Control System (HCS).

Computational Modeling of Bioreactors Based on Spherical Catalyst Particles

R. Baronas[1], J. Kulys[2], L. Petkevičius[1], K. Petrauskas[1]

[1] Institute of Computer Science, Vilnius University
[2] Life Sciences Center, Vilnius University
romas.baronas@mif.vu.lt

A bioreactor is the heart of many biotechnological systems that is used in agricultural, environmental, industrial, medical and other applications. Miniaturized biotechnological systems have gained great attention, and various microreactors with immobilized enzymes have been developed and applied in recent years. The bioreactors are usually designed and operated to provide the environment for the product formation.

In this work, batch and continuous stirred tank reactors based on an array of spherical microbioreactors were mathematically modeled by two-compartment models based on transient reaction-diffusion equations containing a non-linear term related to the Michaelis-Menten kinetics of the enzymatic reaction with addition of the mass transfer of the substrate outside the catalyst region.

The influence of the physical and kinetic parameters of the microbioreactors on the transient effectiveness of the bioreactor system and on the process duration were computationally and partially analytically analysed at transition and steady state conditions in a wide range of model parameters. Effective configurations of microbioreactors have been determined. The computational simulation was carried out using the finite difference technique. The simulation results showed non-monotonic effects of the transient effectiveness and nonlinear effects of the internal and external diffusion limitations as well as adsorption capacity of the microreactors on the transient effectiveness.

This research was funded by a grant (No. S-MIP-17-98) from the Research Council of Lithuania.

Examination and Classification of Data in Digital Forensics

S. Bhandari, V. Jusas

Department of Software Engineering
Kaunas University of Technology
sandeepak525@gmail.com

Digital forensics is a branch of forensic science encompassing the recovery and investigation of material found in digital devices, often in relation to computer crime. The digital investigation process can be categorized into four different phases as a collection, examination, analysis, and reporting. In the research work, two different command based digital forensic tools namely Log2timeline and Psort are used to extract and collect the data from timestamps from various system files found on a typical computer. The extracted data available in two different formats namely XLSX AND L2TCSV having 7 and 17 different fields, respectively. The main aim of both formats is the same, but difference is only in the level of detail. For better examination and extraction of relevant information, a new research methodology is developed. The research methodology is split into four abstraction level of events and artefacts according to a timeline, namely Events: High Level (New entries and Web Surfing), Events: Low Level (Web Surfing and actions of modifying), Artefact Location: High Level and Artefact Location: Low Level. The main aim of the new methodology is to increase the understandability and visualization of data regardless volume and heterogeneity of data to reduce time for collection of digital evidence. In the future research, an ontology-based technique will be developed for better and faster visualization of data for digital forensics.

Cryptocurrency Technologies (Blockchain)

R. Bieliauskas

CTO of CoinGate
JSC "Virtualios valiutos"
rytis@coingate.com

Cryptocurrencies are based on a new revolutionary technology called Blockchain. As defined by the Oxford dictionary, Blockchain is "a system in which a record of transactions made in Bitcoin or another cryptocurrency are maintained across several computers that are linked in a peer-to-peer network."

Blockchain is different from traditional computer systems as its security model is not based on hierarchy and limiting access to information or permissions, but on cryptography and mathematical calculations. It allows creating an open-source decentralized platform / protocol like Bitcoin, which is borderless, decentralized, transparent, censorship-resistant, and permissionless. Just like the open Internet enabled not only permissionless access, but also permissionless innovation in Information Technologies, Bitcoin enables permissionless access and innovation in Financial Technologies.

System Architecture of the Fatigue and Workability Evaluation Platform

L. Bikulčienė, T. Blažauskas, E. Butkevičiūtė, A. Muliuolis

Kaunas University of Technology
liepa.bikulciene@ktu.lt

Human fatigue is the result of prolonged or unusual work manifesting as attention and work capacity reduction for a certain period of time. Statistical data shows that 20% of population is suffering from fatigue, which reduces the work capacity and quality of life. In the presentation the information about Eureka project Fatigue – "Non-intrusive human fatigue assessment" will be introduced. The aim of this project (which involves 6 partners) is to develop platform with complex passive multi-level fatigue monitoring system and workability evaluation system designed in order to provide an integrated service in professional safety and health area. This monitoring and decision-making systems for everyday use with the decision and reasoning in real time across multiple application areas, such as the rehabilitation, sport and workplaces. Lithuanian partners will develop fatigue monitoring system, based on use of unobtrusive sensors, interactive questionnaires and vital signal analysis during daily activities. Integrated solution of different sensors, smart interfaces, modelling and data analysis techniques should warrant that the created system will be comfortable and effective for assessment individuality and dynamics of fatigue level and workability state. During this project KTU researchers will develop data mining methodologies and algorithms (anomaly detection, prediction, decision making, data integration and etc.) for monitoring of vital signs. They will also be involved in analysis of stochastic processes and health characteristics, information from different sensors as well as analysis and integration of information in order to provide the algorithms for effective and robust decision making and creation of software modules for data processing and feedback, i.e. providing recommendations for different use cases.

Multivariate Permutation Test for a Randomized Controlled Trial in the Presence of Small Sample Sizes

S. Bonnini

Department of Economics and Management
University of Ferrara, Italy
bnnsfn@unife.it

The work concerns the application of a multivariate test in a randomized controlled trial with two independent groups of patients (treated group and control group).

The goal of this study consists in testing the effect of a new physiotherapeutic treatment, based on a myotensive and proprioceptive technique, for the stimulation of the muscle spindles. Specifically, the technique was applied to the intrinsic tonic foot muscles.

A sample of seven healthy subjects, football players from 20 to 30 years old, were treated with the myotensive technique under study. On another group of seven similar subjects, the application of the treatment was only simulated. This second sample was considered as a control group.

To measure the treatment effect, rasterstereography and baropodometry were applied. The former consists of projecting a grid of horizontal light lines on the subject's back through laser pointers in order to reproduce a three-dimensional image of the back. A suitable software processes the projected grid and automatically produces the three-dimensional image of the back and identifies the main anatomical landmarks.

A total of 85 numeric response variables, concerning static and dynamic rasterstereography and baropodometry, were taken into account and the differences between observed values after the treatment and before the treatment were computed. A two-sample multivariate test must be applied to these differences (by comparing treated and control group), in order to determine the efficacy of the treatment.

Given the high number of variables and the small sample sizes, the only possible solution can be found within the family of combined permutation tests. This solution guarantees power, flexibility and robustness of the test because it does not require that the underlying family of multivariate distributions is known. Thus the methodology is valid and reliable without being affected by the plausibility of stringent distributional assumptions.

Narrative Detection for Lithuanian Language

M. Briedienė[1], T. Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Baltic Institute of Advanced Technology
monika.briediene@vdu.lt

Automatic narrative detection is an important tool in media analysis, e.g., propaganda identification, elections or marketing campaign tracking, etc. However, it is very hard to detect, i.e. it is easy to miss relating elements of narrative while it is too late to be of interest, or unintentionally to assign false positives to make the case look stronger. Hence, current research results show number of promising directions, but most of them are still at the early stage of research. Almost no results are reported for Lithuanian language, except early ideas in [1]. Moreover, we were not able to find anything on multi-language analysis. However, fast evolution of Machine Learning, and especially, Deep Learning, based methods and easier almost real-time access to different medias sources, looks promising for a number of complex language technologies related applications, including narrative detection. Based on these assumption, we present a study on automatic detection of narrative structure for textual sources in Lithuanian language using automatic machine learning methods. Constituent analysis of a type where narremes are considered to be the basic units of narrative structure could fall within the areas of computer linguistics, natural language processing, semiotics.

This research is novel and challenging due to the following reasons: 1) In order to obtain meaningful results, it is necessary to take a large amount of data; 2) Lithuanian language is morphologically complicated.

The main contribution of this work is study of narrative structure and the ways that it affects personal perception. While in principle the word may refer to any systematic study of narrative, in practice its usage is rather more restricted. A narrative is a report of connected events, real or imaginary, presented in a sequence (the chain) of written or spoken words, or graphical elements, too. Labov and Waletzky [2] defined a structure of narrative consisting of three elements: the orientation, the

complicating action, and the evaluation. Our goal is, for every clause in a narrative, to label it with one of the elements of narrative structure.

We analyze a fundamental problem: how to choose automatic methods that could achieve the highest accuracy in our solving narrative detection task. The related research analysis will help us to select the methods which have demonstrated the best results on the other languages and apply them to the Lithuanian corpora. We look forward to examining the relationship between narrative structure and discourses by examining whether the first determination of the structural element of each sentence fragment can help to detect indirect discourse relationships.

[1] Mandravickaitė, J., Kalinauskaitė D., Krilavičius, T. Visualization of Narrative Structure. In the 1th International Scientific Conference "Challenges To The National Defence In Contemporary Geopolitical Situation", 2018.

[2] Labov, W, Waletzky, J. Narrative Analysis. University of Washington Press, Seattle, W.A., 1967.

Sublevel Structure Similarity Metric for X-Ray Images Comparison

J. Brusokas, L. Petkevičius

Institute of Computer Science
Vilnius University
jonas.brusokas@mif.vu.lt

The number of collected medical images by hospitals and medical institutions have been increasing drastically in recent years. In order to deal with large amounts of data medical image compression, image reconstruction and lossless data reduction remain open problems. In this work, the metrics for comparing two x-ray type medical images at various scales and intensity regions are developed. Most of the image similarity metrics work well on RGB images, where subsets of various intensity levels are not as important as in medical images. Medical images by radiologists are often analysed at various slices of intensity levels, good image quality within slices are critical in making good decisions. New metrics include structure information as well as distances between the various ranges of levels. The metrics value decreases as the quality of the images decreases. The computation results were evaluated on multiple open access X-ray datasets.

Colloc – A Toolset for Multiword Expressions Identification

I. Bumbulienė[1], J. Mandravickaitė[1], L. Boizou[2], T. Krilavičius[1,2]

[1] Baltic Institute of Advanced Technology
[2] Vytautas Magnus University
tomas.krilavicius@bpti.lt

Multiword Expressions and its automatic identification are rather important topics in Language Technologies due to their omnipresence, complications in translating them and their complex semantics. A number of different methods are used to identify MWE, such as Lexical Association Measures, Machine and Deep Learning as well as various complex approaches. However, the problem is still far from being solved, especially for resource scarce languages as Lithuanian. Currently, DL methods lead to improving results for many NLP tasks including MWEs identification. Hence, we apply a number of different methods for MWE identification in Lithuanian and report a framework and Colloc tool for automatic MWEs identification in Lithuanian.

Framework is based on sequence models – CRF (Conditional Random Fields) and RNN (bi-LSTM, Recurrent Neural Networks) trained using GloVe embeddings (331 thousand unique words, dimension of 200) and part-of-speech as features. Due to usage of sequence models, the tool can identify MWEs of different length as well as MWEs with insertions. F1-Score the CRF and RNN models' combination in the tagger reaches for 49% for MWEs of different length. Moreover, Colloc allows to query the database of MWEs, which was filled up with linguists approved automatically extracted MWEs. The tool is available on-line, see www.mwe.lt for more information.

Benchmarking Secondary Schools Using Students' Results at University

A. S. Camanho[1], M. C. Silva[2], F. Barbosa[1]

[1] Faculty of Engineering of the University of Porto, Portugal
[2] CEGE, Católica Porto Business School, Portugal
acamanho@fe.up.pt

This research uses data on university students' first-year scores as a means to benchmark secondary schools on their ability to lead students to university success. Data from two universities in Portugal, the University of Porto (with 14 faculties) and the Portuguese Catholic University (with 7 faculties), are used to compute several indicators, based on which secondary schools are compared. The performance of more than 10.000 students from 65 different degrees was explored for a three year period. There are mainly two research questions addressed in this paper: (1) what are the determinants of success for first-year university students? and (2) what is the relative performance of secondary schools in terms of students' attainment in higher education? These research questions were explored using Regression and Data Envelopment Analysis (DEA) models.

Regression analysis was used to identify the determinants of success for first-year students, assuming that students' performance can depend on student characteristics (e.g., grades on entry to higher education and gender) and secondary school characteristics (e.g., private vs. public ownership). University success was measured through two indicators: number of ECTS completed by first-year students and the final grade at the end of the first year (this final grade is normalised per degree attended and cohort to allow comparability between different degrees and years). Conclusions from the preliminary analysis point to the importance of students' grades on entry (based on national exams at the end of secondary education). The type of school appears as the second most important factor in determining student's success, with public schools performing better than private. In relation to gender, female students tend to have better performance than males in the first year of higher education.

Overall, these factors are more important to explain the average grade obtained at the end of the first year than the number of ECTS completed by the students. We also found some variability across degrees in terms of the impact of these factors on students' performance in higher education.

Concerning the benchmarking of secondary schools based on students' attainment in higher education, a DEA model was constructed to compare secondary schools on the extent to which they prepare students to university success. For this purpose, a set of outputs relating to success in the first year of higher education was defined (e.g., percentage of students with top performance in the degree attended, average number of ECTS completed by the students, and average score of students at the end of the first year, normalised according to the degree attended). We developed a composite indicator model, formulated with a Directional Distance Function, that allows conducting benchmarking evaluations in the presence of negative data. Note that in the context of our study, negative data occurs due to the normalization of students' first year scores within their degrees, to allow comparability across different degrees. Our approach to schools' benchmarking is unprecedented in the literature, as it evaluates secondary schools based on the performance of students in the next educational stage.

We found that schools' ranking based on their ability to prepare students for university success provides a very different picture from schools' rankings based on results on national exams at the end of secondary education. Given these findings, we propose complementing schools' performance assessments with indicators that account for the preparation of students for success in future challenges, which is undisputedly a key objective of secondary education.

Using Profiling to Assemble Freelance Software Development Teams in the Context of GDPR

M. L. Despa[1], I. Ivan[1], E. N. Budacu[1], M. Visan[2]

[1] Bucharest University of Economic Studies, Romania
[2] Department of Engineering, Mechanics, Computers, Romanian Academy, School of Advanced Studies of the Romanian Academy (SCOSAAR)
mihai.despa@gdm.ro

Assembling software development freelancer comprised teams is a challenging task for a project manager. He needs to choose individuals that would work well together and have the skills to match project complexity. Considering the aim of agile methods to organize cross-functional teams consisting of multi-disciplinary individuals we evaluated the features and challenges of forming them. This is a lengthy endeavor which is prone to human bias. It also has considerable restrictions imposed by the sheer volume of information a human decision maker can process and analyze. Computerized data analysis on the other hand is free of bias and it can process large data sets very fast. For the current research process data was collected automatically from freelance platforms that allow access to developers' public profiles. The targeted platforms make available information about freelancer's skills, experience, background and education. By using crawling and scrapping technics a large dataset containing profile information was automatically compiled. Data was validated, ensuring it's correct and relevant. Data was normalized to ensure it is easily quantifiable and can be measured using a common scale. Selection criteria for assembling the team was determined automatically using statistical methods. Based on the identified criteria a template was built to match the profile of the ideal team member. Machine learning techniques were used to group team members that would work together effectively. An indicator was built to benchmark team member's profile against the template and determine the highest ranking candidates. The entire process was analyzed from the General Data Protection Regulation (GDPR) perspective to ensure data handling is fully compliant with the latest regulation. Conclusions were formulated regarding the topic of using profiling to assemble agile freelance software development teams. Research process limitations were enunciated and future research topics were submitted for debate.

The Laboratory of an Education Economist. Testing Cures for Disadvantaged Students

K. De Witte

Faculty of Economics and Business at KU Leuven, Belgium
kristof.dewitte@kuleuven.be

Kristof De Witte overviews his research agenda on socio-economic segregation. In particular, he discusses various quasi-experimental evaluations of interventions to change the odds for disadvantaged students. On the one hand, he presents the effects of additional resources at school level on cognitive and non-cognitive outcomes and illustrates that that extra resources at the municipality level only result in grade inflation. On the other hand, he argues that information shocks provided by making school quality information public changes the socio-economic composition of schools. He shows that in the longer run, these findings are alarming as the inability to break the vicious circle for low SES students leads to more school dropout, and that particularly those students have a lower return to education.

Performance Evaluations for Supervised Classifiers of Gaussian Random Field Observations

K. Ducinskas[1,2], L. Dreiziene[1]

[1] Klaipėda University
[2] Vilnius University
k.ducinskas@gmail.com

The problem of classifying a scalar Gaussian random field observation into one of two populations specified by different parametric mean models and a parametric spatial covariance function is considered. Authors concerns with classification procedures associated with Bayes Discriminant Function (BDF) under deterministic spatial sampling design. In the case of parametric uncertainty, the ML estimators of unknown parameters are plugged into the BDF. The probability of correct classification is considered as a performance measure of proposed classifier. Closed-form expressions of the actual probability of correct classification (APCR) associated with aforementioned plug-in BDF are derived both for Geostatistical and Markov Gaussian models of spatial data. Approximations of the expectation of ACPR is derived by using ML estimator properties under increasing domain framework. Numerical analysis of proposed classifiers performances are done with simulated and real data.

GeoR package of statistical software R is used in simulation of the Gaussian Random field realization. Stationary geometrically anisotropic Gaussian random field with exponential covariance function sampled on regular 2-dimensional lattice is used for illustrative examples. Different spatial sampling designs are compared.

Various types of spatial data models for invasive species (zebra mussels) distributed in the Curonian Lagoon are considered and compared by proposed performance measure. Advanced models are proposed to the mapping of presence and absence of zebra mussels in the Curonian Lagoon.

An Approach for Networking of Wireless Sensors and Embedded Systems Applied for Monitoring of Environment Data

D. Dzemydienė, V. Radzevičius

Institute of Data Science and Digital Technologies
Vilnius University
dale.dzemydiene@mii.vu.lt

Our research area concerns the possibilities of wireless sensor networks and embedded systems, which are the core components of the architecture of Internet of Things (IoT). Mechanisms of such interconnected devices have to work by properly presented goals. Developing modern embedded systems requires intellectualization methods based on certain artificial intelligence methods such as machine learning, neural networks, decision trees, and rules for goal-oriented control. The capabilities of such systems are based on the technology of the IoT, but the capacities of the interconnected embedded systems are quite limited. The device control systems that monitor environment settings are connected to the network. They are becoming more and more able to control other devices based on their real-time values. Operational management systems are usually managed by the principles of monitoring of live values, but the identification of situations and control actions have to achieve requirements of very quick reactions. However, the small capacities of embedded systems restrict us from creating large data repositories and knowledge bases inside of these systems for assessing situations. Therefore, it is necessary to address the issues of network allocation and system resource allocation and the right architecture offer. In order to ensure the integration of sensors into a common management system, this work proposes a layer of interconnected sensors and controller's network applied on the base of other standardized network layers. The prototype method is used for evaluation of low-performance embedded systems and wireless sensor networks (WSNs), and enable to collect the data of environment parameters for operative control needs.

DSS – An Evolving Class of Information Systems

F. G. Filip

Section of Information Science and Technology
Romanian Academy
ffilip@acad.ro

The talk starts with the presentation of basic concepts of a particular class of information systems, namely DSS (Decision Support Systems) which are meant to help the decision-maker to solve complex decision problems that count. Various classifications made by specific criteria such as the type of support, number of users, decision-maker type and influence level, technological orientation, are described. The evolution trends are presented with a particular emphasis on modern I&C (Information and Communication) technologies utilized and business models adopted. A set o design criteria and the associated multi-participant decision making the practical method to choose an adequate solution are discussed.

Reducing Gender Bias in Data for Lithuanian

V. Fomin, D. Amilevičius

Vytautas Magnus University
darius.amilevicius@vdu.lt

As application of artificial intelligence in different domains becomes more ubiquitous, scholars raise questions on ethical considerations with regard to privacy, decision bias, algorithmic transparency and accountability. Theoretically, machines are supposed to deliver unbiased decisions. Recent examples, however, show the contrary – even algorithmic mind can be prejudiced. One typical example of bias in AI is that of gender in language translations. To assure AI functionality delivers bias-free results, the underlying machine learning process must be properly managed from input to output – including data, algorithms, models, training, testing and predictions – to make sure that bias is not perpetuated. Speakers of grammatically genderless language may have impression that it is superior compared to grammatical gender language, as they supposedly genderless character is seen as an expression of gender equality. But a lack of grammatical gender does not automatically reflect a more gender-neutral society. Due to the absence of grammatical gender femaleness and maleness can only be expressed through lexical and socially gendered forms. Morphologically gender-neutral nouns often carry a hidden cultural or social gender bias. In Lithuanian, a male noun has two referential functions: a male-specific and generic function. For this reason, the former case goes close to bias in genderless languages. In our research we detected the presence of bias in Lithuanian text which links women to less prestigious jobs, while men – to more prestigious ones. The problem is not related to the morphology. We used word embedding (word2vec and FastText) and debiased word embedding methods. The result is that debiased word embedding method reducing gender bias is more effective than bias fine-tuning method. The gender biases in the embedding could capture useful statistics. However, given the potential risk of having machine learning algorithms that amplify gender discrimination, we should use the debiased embedding s as much as possible.

This work is supported by the grant No. 09.3.3-LMT-K-712-01-0173.

High Performance Computing: Platforms and Techniques

E. M. Garzón

University of Almeria, Spain
gmartin@ual.es

It is of utmost importance to overcome the challenges of the High Performance Computing (HPC) to advance in the main topics of the modern computation, such as the Internet of things, Data Mining, Artificial Intelligence and so on since they have tremendous computational requirements. Supercomputers provide great computational resources, which can be harnessed by Cloud technology.

Moreover, thanks to technological and architectural advances, current HPC platforms are not only located in Supercomputing Centers, but also standalone computers, smartphones, embedded systems can be considered as HPC platforms if they are appropriately exploited. These architectures are based on heterogeneous multi-core processors, and they include additional resources to take advantage of different kinds of parallelism in applications. This tutorial revises the keys of the current HPC platforms from the point of view of the hardware technology and software interfaces, with a special focus on the representative HPC applications.

Reviewing MDD from the Causal Modeling Perspective

S. Gudas, A. Valatavičius

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

The model driven development (MDD / MDA) of the software systems (cyber social systems, cyber enterprise systems) is now being improved by enhancing the modeling methods on CIM / PIM / PSM layers and model transformation techniques.

Evaluating the MDD techniques from the system analysis perspective, most methods, modeling languages, and frameworks (BPMN, DMN, UML, SySML, MODAF, UPDM, UAF) capture the results of external observation. That is, they belong to black box / grey box modeling paradigms.

In parallel, viable results today are achieved with the development of smart systems, autonomic systems and other types of the cyber-physical systems (CPS). A cyber-physical systems (CPS) engineering is based on the causal dependencies of the subject domain. That is, CPS development is on the white box / grey box paradigm.

Therefore, the proposed MDA transition to the white box approach, based on causal modeling of the domain, is reasonable.

The methodological background of our approach to the enterprise software systems engineering is shifting to the white box paradigm by integration of traditional MDA schema with the causal modeling of enterprise domain. The modified MDA schema includes the new layer of the domain knowledge discovery, and frameworks for enterprise causal dependencies modeling. The peculiarity of the modified MDA is a focus on the cross-layer transferring of domain causality using the internal model principle.

The presented frameworks will help to trace the domain causal dependencies across the layers of the software system development, and to determine the influence of domain causality to the integrity and interoperability of the application.

Gamification for Engaging IT Students in the ERP Course: A Case Study

M. Heričko, A. Rajšp, T. Beranič

Faculty of Electrical Engineering and Computer Science
University of Maribor, Slovenia
marjan.hericko@um.si

Gamification is the use of game design elements in non-game settings to engage participants. When applied in Higher Education, its success depends on the ability to involve students, since students' engagement is correlated positively to their success, satisfaction, and academic achievements.

ERP systems are huge and complex, therefore, their introduction into IT&SE Study Programs presents a challenging task, especially if students are expected to gain an understanding of ERP systems from both, a functional (business process) and implementation perspective.

In this presentation, the results will be given of a two-year study aimed at investigating the appropriateness and effectiveness of the business simulation game approach for introductory ERP classes. The results show positive effects on the students' engagement towards the gamified learning. Based on the positive student and instructor feedback, we can conclude that business simulation games are the right approach when introducing ERP systems such as SAP within an ERP course. Furthermore, we also measured the perceived ERP system usability and detected the very positive effect of gamification.

Exact Versus Evolutionary Methods to Derive Pareto Fronts in the Mean-Variance Portfolio Investment Problem – The Evidence from the JKMP Data Set

P. Juszczuk[1], I. Kaliszewski[2], J. Miroforidis[2], D. Podkopaev[2]

[1] University of Economics in Katowice, Poland
[2] Systems Research Institute, Polish Academy of Sciences
janusz.miroforidis@ibspan.waw.pl

In order to really grasp trade-off between return and risk in the set of efficient portfolios of the Markowitz mean-variance investment problem, this set or its approximation is to be derived. We investigate exact and evolutionary methods of deriving mean-variance Pareto fronts with our own large-scale JKMP test problems. We also show how to navigate over the Pareto front with the use of so-called compromise half lines to unveil the decision-maker's preferences.

Low-Cardinality Approximations of the Pareto Front in the Mean-Variance Portfolio Investment Problem with a Non-Risky Asset

P. Juszczuk[1], I. Kaliszewski[2], J. Miroforidis[2], D. Podkopaev[2]

[1] University of Economics in Katowice, Poland
[2] Systems Research Institute, Polish Academy of Sciences
podkop@ibspan.waw.pl

Keeping the small number of assets in the portfolio while maximizing the mean return and minimizing the risk is one of the most important issues discussed in modern portfolio theory. Convex combinations of assets, as well as convex combinations of portfolios, allow keeping these numbers low. On the other hand, the quality of a lower approximation of the Pareto front observed in the region of lower values of mean return is not satisfactory. We discuss the case where a non-risky asset, i.e. a minimum-mean-minimum-variance instrument exists and it is included in the convex combinations. We show that including the decision maker's preferences in the model and representing them as a compromise half-line can be used to decrease the computational complexity of the method, as only a part of the lower approximation of the Pareto front needs to be derived.

All presented experiments are conducted with the use of newly generated datasets based on the original Beasley OR library format.

Algorithmic Thinking Through Computational Making

A. Juškevičienė

Institute of Data Science and Digital Technologies
Vilnius University
anita.juskeviciene@mii.vu.lt

Algorithmic thinking is the main component of computational thinking (CT). CT is in line with many 21st century skills necessary for digital learners. However, it is still a challenge for educators to taught CT in an attractive way for learners, also to find support to curriculum design and assessment. To address this problem, the literature review on CT in education was conducted and the main ideas of CT assessment were identified. The results show that modern technologies are widely used for learning enhancement and algorithmic thinking improvement. The implications of these results is that modern technologies can facilitate effective learning and CT skills gaining.

Decisions in Human Centric Multiagent Systems: Rationality and Some Cognitive Biases

J. Kacprzyk

Systems Research Institute
Polish Academy of Sciences
kacprzyk@ibspan.waw.pl

We are concerned with systems in which there are multiple agents, humans or software agents imitating the human judgments, behavior, perception, evaluations, etc. The agents are faced with a set of options (alternatives or courses of action) and should find an option or a set of options that can be considered as the best acceptable by the entire group. The agents provide their testimonies as to their opinions on the goodness of options which can be given as preference relations, utility (valuation) functions, sets of approved/disapproved options, etc. The best option is then found on the basis of these testimonies and also some other aspects like attitudes of the agents. The process is then assumed to be run with the help of a moderator, a "super-agent" who is guiding the process. First of all, the process is meant to involve a consensus reaching step in which the source, usually highly different testimonies of the agents are made closer, i.e., the agents change their testimonies usually persuaded by the moderator. Then, after reaching a "consensus," i.e., closer testimonies, some group decision making (or voting) tools are used to obtain solutions which are mainly some cores, i.e., sets of options that are undominated with respect to other options in the opinion of most agents. We assume quite a general approach with fuzzy preferences and a fuzzy majority. It is shown that such an approach exhibit, first, some properties of a greedy, or selfish, approach since the agents are just taken into account their testimonies. A new approach is proposed in which some fairness occurs in the sense that testimonies of all agents are accounted for.

Moreover, as this fairness oriented approach involves changes of the agents' testimonies, with a fair suggestion of such changes to all agents,

a new model is proposed which takes into account some well known cognitive bias, the status quo bias. The cognitive bias is meant in the sense of Kahneman and Tversky, i.e., that people make judgments or decisions in the ways that are systematically different than obtained from traditional economic models, and the cognitive bias does not necessarily lead to bad decisions. Among the cognitive biases the status quo bias, i.e., that people tend to avoid larger changes is shown to be promising in the context considered. Possible use of some other cognitive biases is mentioned.

Analysis of Lombard Speech Using Parameterization and the Objective Quality Indicators in Noise Conditions

K. Kąkol[1], G. Korvel[2], B. Kostek[1]

[1] Audio Acoustics Laboratory
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology, Poland
[2] Institute of Data Science and Digital Technologies
Vilnius University
grazina.korvel@mii.vu.lt

The aim of the work is to analyze Lombard speech effect in recordings and then modify the speech signal in order to obtain an increase in the improvement of objective speech quality indicators after mixing the useful signal with noise or with an interfering signal. The modifications made to the signal are based on the characteristics of the Lombard speech, and in particular on the effect of increasing the fundamental frequency F0. The recording session includes sets of words and sentences in Polish, recorded in silence, as well as in the presence of interfering signals, i.e. pink noise and so-called bustle (called babble speech), also referred to as the "cocktail-party" effect. Research on the Lombard speech often focuses on subjective studies of speech intelligibility. There are, however, objective indicators such as PESQ (Perceptual Evaluation of Speech Quality) and P.563, which are used in studies of quality of telecommunication channels. The study shows that increasing the fundamental frequency results in increased values of the speech quality index, measured using the PESQ (Perceptual Evaluation of Speech Quality) standard. The research carried out consists of several stages: (1) recording speech samples (words and sentences) without and in the presence of pink noise and babble speech (the so-called cocktail party effect), i.e. the reference signal ("clean" speech), and then recording the same words/sentences in the presence of additional disturbances forcing the Lombard effect in speech recordings to occur; (2) analyzing differences between "clean" speech and the Lombard speech based on objective audio parameters; (3) mixing speech recordings with pink noise with a different signal to

noise ratio (SNR) in order to measure PESQ MOS coefficients; (4) measuring the PESQ coefficients of the reference files ("clean speech") that are processed by increasing the F0 value and sound intensity level, and then the same files mixed with pink noise and babble speech interfering signals; (5) repeating step (2), i.e. analyzing the difference in objective parameters and indicating whether these differences are statistically significant.

From Manual to Automated Measurement of Information Density

D. Kalinauskaitė[1,2]

[1] Baltic Institute of Advanced Technology
[2] Vytautas Magnus University
danguole.kalinauskaite@bpti.lt

Determining information density of texts is a big challenge in natural language processing. Recent developments in this field have spawned a number of solutions to evaluate information density. Nevertheless, a shortfall of most of these solutions is their dependency on the genre and domain of the text. Different text genres and/or registers have their own distinctive structure and vocabulary use.

The notion of information is only formal here, i.e. information is defined as semantic, pragmatic, and only measurable in relative terms. A definition of information density is elaborated involving informativity (a relative measure of semantic and pragmatic information) per clause. Numerous experiments have related information density to readability, memory, quality of students' writing, aging, and prediction of Alzheimer's disease.

This research proposes a methodology for automatic detection of information-dense texts. The methodology is based on the relationship between information density and the use of certain lexical and syntactic features. This part of the research discusses experiments with two corpora, compiled from the research papers and their abstracts, as well as experiments with different features of texts, selected for the analysis. The results, presented here, are part of the whole methodology proposed for automatic analysis of information density. While information density is seen as too complex to measure globally, a study of both lexical and syntactic features allows a comparison of information density between different text genres.

Control of Emotion as Reaction to a Dynamic Virtual 3D Face: A Comparison of the Predictor-Based Control Schemes

V. Kaminskas, E. Ščiglinskas

Vytautas Magnus University
vytautas.kaminskas@vdu.lt

This paper introduces the application of predictor-based control with constraints of human emotions as reaction to a dynamic virtual 3D face. A dynamic woman 3D face is observed in virtual reality or without virtual reality. We use changing distance-between-eyes in a woman 3D face as a stimulus – control signal. Human response to the stimulus is observed using EEG-based excitement signal – output signal. The technique of dynamic systems identification which ensure stability and possible higher gain of the model for building a predictive input-output model of control plant is applied. Three predictor-based control schemes with a minimum variance or a generalized minimum variance control quality and constrained control signal magnitude and change rate are developed. High prediction accuracy and control quality are demonstrated by modelling results.

A Decentralized System for Managing Micro-Credentials Based on Blockchain 2.0

A. Kamišalić, M. Turkanović, M. Heričko

Institute of Informatics
Faculty of Electrical Engineering and Computer Science
University of Maribor, Slovenia
aida.kamisalic@um.si

Blockchain technology enables the creation of a decentralized and distributed environment in which transactions and data are stored in a publicly available ledger not controlled by a central authority. Transactions are secure, ubiquitous and trustworthy as they employ cryptographic principles.

We will present a global decentralized blockchain-based platform and ecosystem called EduCTX. The platform offers a comprehensive and unified digital environment for managing micro-credentials for individuals by educational institutions as well as other potential stakeholders such as companies, institutions, and organizations. Furthermore, by enabling public access to the ledger, the platform provides organizations the possibility to develop their digital services in order to automate the evaluation of individuals' skills and knowledge.

The EduCTX platform is implemented as a consortium-based P2P network between the aforementioned stakeholders, using the Ethereum blockchain platform. The Ethereum platform enables the creation of decentralized applications based on smart contracts, which is the key functionality of Blockchain 2.0 platforms. Smart contracts are programmable digital protocols, which are automatically executed based on the embedded logic. The platform facilitates the anonymous storage of personal data, thus addressing privacy issues.

The developed platform is the basis of the EduCTX initiative, which aims at bringing together different geographically distributed stakeholders to create an effective, simplified and ubiquitous digital environment to avoid linguistic and administrative barriers. The platform contributes

to the modernization of processes and supports the development and deployment of innovative digital services. Using EduCTX can shorten the time-consuming, costly and burdensome processes of organizations. Another important aspect of the platform lies in its social importance as it enables individuals to have the equal possibility to share their competences with potential employers.

Classification of Business Documents with Neural Network

S. Karakatič, M. Heričko, L. Pavlič

University of Maribor, Slovenia
luka.pavlic@um.si

Manual document classification, in general, is very time consuming and error prone activity. Therefore, it is a perfect candidate for an automation. The document classification problem has already been widely researched. However, in research community we are missing so far a research focus on mid-length documents, classified in one of the closely related classes with low variance.

This is why we present a novel approach to the classification of documents in the business investment domain. We employ machine learning technique heavily in it. Our proposed system uses the vector representation of words (word2vec) and whole documents (doc2vec) which are later used as an input into the document classification system. A convolutional neural network is used to classify documents in one of the predefined classes, e.g., “not interesting”, “lead”, “investment”. The convolution layer of the networks traverses the whole document with the variable size of the convolution layer window.

Lack of data would normally affect the neural network model training. With transfer learning from existing word2vec models we managed to overcome this obstacle. The classification accuracy and F-score metrics are used to determine the quality of the proposed system.

During the presentation we are demonstrating techniques used in our novel approach to document classification. We are presenting web-based prototype implementation also.

Expected Error Regret in Linear Discrimination of Balanced Spatial Gaussian Time Series

M. Karaliūtė[1], K. Dučinskas[1,2], L. Šaltytė-Vaisiauskė[2]

[1] Vilnius University
[2] Klaipėda University
mkaraliute@gmail.com

The problems of discriminant analysis of spatially correlated Gaussian data were intensively considered previously (see Switzer (1980), McLachlan (2004), Saltyte-Benth and Ducinskas (2005)). In these papers, theoretical results were derived under the assumption of statistical independence between observation to be classified and training sample. In the present paper, we avoid this tough restriction. The problem of supervised classifying of the spatial Gaussian time series (SGTS) observation into one of two populations specified by different regression mean models and by common non separable covariance function is considered. In the case of complete parametric certainty and with the fixed training sample locations, the formula of conditional Bayes error rate is derived. In the case of unknown regression parameters and temporal covariance matrix, their ML estimators are plugged into the Bayes discriminant function. The actual error rate with plug – in Bayes discriminant function derived. These results are multivariate generalizations of previous ones. Numerical analysis of the derived formulas is implemented for the SGTS observations at locations belonging to the 2-dimensional lattice with unit spacing and isotropic exponential spatial correlation. Temporal dependence described by the AR(p) models for stationary time series. Expected error regret (EERG) as the difference between conditional Bayes rate and the actual error rate is estimated.

Application of Social Network and Content Analysis Methods for Assessing the Dynamics of Facebook Groups

R. Kasperienė[1,2], T. Krilavičius[1,2]

[1] Vytautas Magnus University
[2] Baltic Institute of Advanced Technology
kasperiene@gmail.com

The relationship between the content that is generated by the users of social networks and their dynamics have been analyzed by many scholars. However, due to favorable data policies, the majority of studies have been carried out by analyzing Twitter data. In addition, such research on Facebook (FB) groups (esp. political) is usually qualitative. The present study analyses the dynamics as well as topic dynamics of radical right political groups on FB by employing a quantitative research methodology. The current paper draws on a large data set that is comprised of posts from FB groups. Overall, there are 79 728 posts which are made up of more than 2 million words and were generated within the timespan ranging from 2010 to 2018. The experimental set up compares the general dynamics and the dynamics of activity on four topics in two radical right FB groups (i.e., pro-Russian and other radical right) in Lithuania. The results show that the year 2014 was important for the radical right FB groups in Lithuania. Newly created pro-Russian FB groups started growing rapidly, whereas the posting activity in other radical right FB groups started to decrease. The topic word Lithuania is relevant for the whole activity time when it comes to all the radical right FB groups. Such topic words as Russia and land correlate with national and international political crisis.

Bone Age Assessment Using Deep Learning Detection Models

V. Kazlauskas[1], A. Neverauskienė[2], L. Petkevičius[1]

[1] Institute of Computer Science
Vilnius University
[2] Department of Radiology, Children's Hospital
Affiliate of Vilnius University Hospital Santaros Klinikos
linas.petkevicius@mif.vu.lt

The visual assessment of skeletal bone age is a common clinical practice to diagnose disorders in child development. In this work, we describe a fully automated deep learning approach to the problem of bone age assessment using data from the 2017 RSNA Bone Age Challenge organized by Kaggle and Digital Hand Atlas Database System. The datasets for this competition consists of 12600 and 1388 radiological images, respectively. Each left hand x-ray image in this dataset labeled with estimate of bone age and gender of a patient. Our approach utilizes two-stage deep neural network architecture. In stage one we using CNN based detection model to detect important regions of hand. The second stage using convolutional neural network to predict bone age. We further evaluate the performance of the suggested architecture and compare results with state-of-the-art model results. After the validation of the model, user friendly application was created. Program estimate bone age from given x-ray image, and provide most similar results from database. The program was deployed to Department of Radiology, Children's Hospital, Affiliate of Vilnius University Hospital Santaros Klinikos as auxiliary tool for bone age evaluation.

Semibinomial Conditionally Nonlinear Autoregressive Time Series for Data Science

Yu. Kharin, V. Voloshko

Research Institute for Applied Problems
of Mathematics & Informatics Belarusian State University
kharin@bsu.by

We present a new wide class of parsimonious models for discrete time series $x_t \in A = \{0, ..., N\}$ that we call semibinomial conditionally nonlinear autoregressive time series (ρ-CNAR) of order s. Here ρ is some fixed one-parametric family of probability distributions on A, containing all the conditional distributions of x_t depending on s-prehistory $x_{t-1}, ..., x_{t-s}$ via linear combination of m known nonlinear base functions $f_i(x_{t-1}, ..., x_{t-s})$: $i = 1, ..., m$, with unknown coefficients $\{a_i\}$. This model is a generalization of the model for binary time series developed by the authors earlier [1].

We give some basic probabilistic properties of ρ-CNAR model, construct a family of consistent asymptotically normal frequencies-based statistical estimators (FBE) for the model parameter. We find the optimal FBE within constructed family and show that the optimal FBE is efficient and has some computational advantages w.r.t. the maximum likelihood estimator. These advantages are: explicit form; recursive computation under model extension (adding a new base function $f_{m+1}(\bullet)$; less restrictive uniqueness sufficient conditions. We also find the optimal FBE within the subfamily of sparse FBE, that use only some subset of $(N+1)^s$ frequencies. Based on the constructed formula for the asymptotic variance of the optimal sparse FBE, we propose an empirical choice for the sparse subset of frequencies.

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

[1] Kharin, Y. S., Voloshko, V. A., Medved, E. A. Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series. Mathematical Methods of Statistics, 27(2), 2018, p. 103-118.

Large Scale Sentiment Analysis Using NLP Based Feature Extraction Technique and PSOLinearSVM

K. Korovkinas, G. Garšva

Institute of Social Sciences and Applied Informatics
Vilnius University
konstantinas.korovkinas@knf.vu.lt

Sentiment polarity recognition from text is one of the most challenging research area. In this paper are presented two approaches to increase sentiment classification accuracy and speed: natural language processing (NLP) based feature extraction technique from large scale text data arrays for representative dataset creation and heuristic method PSOLinearSVM based on particle swarm optimization for SVM parameter tuning. The new created representative training dataset was used as input for PSOLinearSVM and tested on two existing labeled datasets: the Stanford Twitter sentiment corpus and Amazon customer reviews. The results was compared with our previous research. It was achieved that proposed methods by using them together significantly increase sentiment recognition accuracy and classification speed.

Investigating Feature Spaces for Isolated Word Recognition

G. Korvel[1], G. Tamulevičius[1], P. Treigys[1],
J. Bernatavičienė[1], B. Kostek[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Audio Acoustics Laboratory
Faculty of Electronics, Telecommunications and Informatics
Gdańsk University of Technology, Poland
grazina.korvel@mii.vu.lt

Much attention is given by researchers to the speech processing task in automatic speech recognition (ASR) over the past decades. The study addresses the issue related to the investigation of the appropriateness of a two-dimensional representation of speech feature spaces for speech recognition tasks based on deep learning techniques. The approach combines Convolutional Neural Networks (CNNs) and time-frequency signal representation converted to the investigative feature spaces. In particular, fractal dimension features of the signal were chosen for the time domain, and two feature spaces were investigated for the frequency domain, namely: frequency tracks obtained from the frequencies and amplitudes of the detected spectral peaks and the modified chromagrams. Both are constructed from a series of short-time Fourier transforms, which were computed along the window speech signal in the time domain. Due to the fact that deep learning requires a sufficiently large training set as the size of the corpus may significantly influence the outcome, thus for the data augmentation purpose, the created dataset was extended by adding various noise levels and mixed with the speech signal. In order to evaluate the applicability of implemented feature spaces for isolated word recognition task, three experiments were conducted: a 10-word, a 70-word, and a 111-word cases were analyzed.

Three Level Parallelisation Scheme for Optimisation Problems Involving Simultaneous Calculations of Multiple Differential Equations

R. Kriauzienė[1,2], A. Bugajev[2], R. Čiegis[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Vilnius Gediminas Technical University
kriauziene@gmail.com

We propose a general methodology for solving optimisation problems when there is a big number of processes available. In this research we investigate a three-level parallelisation algorithm for optimisation problems, different parallelisation levels create different challenges.

At the first level of parallelisation we assume that there exist parallel alternatives to the original sequential modelling algorithm. The first level of parallelisation becomes a part of a new parallel algorithm and the degree of parallelism can be selected dynamically during the computations. The parallelisation speed-up on the first level is not linear, it can lower the efficiency of the whole parallelisation. In this paper as an example we consider the parallelised simplex downhill method.

On the second level, a set of computational tasks with different computational sizes is defined. The work amount distribution between tasks is non-uniform – this makes the parallelisation challenging. This leads to necessity of third level, because a proper load balancing must be performed. As an example we investigate the case when M partial differential equations are solved. The computational sizes of these tasks are non-equal because different discretisation sizes must be used for each equation in order to achieve the same level of errors.

The third level defines parallel algorithms to solve tasks from the second level. As an example we take Wang's algorithm to parallelise the solution of systems of linear equations with tridiagonal matrices. This level can be used alone, however, it is limited due to Amdahl's law.

We presented a general methodology, which combines the parallelisation of a local optimisation algorithm with a standard two level parallelisation. This is a new three-level parallelisation scheme.

Acknowledgments. Computations were performed using resources at the High Performance Computing Centre HPC Sauletekis in Vilnius University Faculty of Physics.

Mathematical Morphology and Multivariate Analysis Based Protein Expression Signal Analysis in Microarray Imaging

A. Kriščiukaitis[1,2], R. Urbanavičiūtė[2], R. Petrolis[1,2], D. Skiriutė[2]

[1] Dept. Physics, Mathematics and Biophysics
Neuroscience Institute
Lithuanian University of Health Sciences
[2] Neuroscience Institute
Lithuanian University of Health Sciences
algimantas.krisciukaitis@lsmuni.lt

High throughput microarray analysis has great potential in wide range of research areas such as protein expression and/or genome analysis. It could be also used in particular disease diagnosis. By putting DNA in an array on some membrane, one can create an array of spots, where each spot contains thousands of identical molecules, consisting of DNA, cDNA or oligonucleotides. Probe molecules, typically labeled with a fluorescent dye, are added to the array. Any reaction between the probe and the immobilized molecule emits a fluorescent signal. Microarray images capture the fluorescence intensity information of these spots reflecting, for example, expression of particular protein or some gene. Technically sophisticated procedures and manipulations during membrane preparation result in burying of valuable information into numerous distortions of microarray images. Fortunately, some distortions do not corrupt this information irreversibly and it could be restored by means of special image processing. So, elaboration of such advanced imaging methods is becoming a key step in development of successful high-throughput microarray analysis.

Mathematical morphology-based methods were used for preprocessing of microarray images, where cancellation of uneven illumination together with shape distortions were performed to standardize analyzed images. Principal component analysis-based methods were used for enhancement of fluorescence intensity estimation precision and to reveal proteomic profiles related with particular types of brain tumors.

Deep Learning for Understanding Human Vision

J. Kubilius[1,2]

[1] Brain and Cognition
KU Leuven, Belgium
[2] McGovern Institute for Brain Research
Massachusetts Institute of Technology, USA
jonas.kubilius@kuleuven.be

How do we recognize what we see? Despite the deceptive ease of perceiving things, explaining how we see turns out to be a supremely difficult task. Only recently advances in computer vision finally brought a class of models, known as deep neural nets, that are capable of matching human and non-human primate performance in several visual perception tasks. Our present aim is to develop these artificial systems further so that they would simultenously (i) predict primate neural and behavioral responses during visual object recognition tasks, (ii) map well onto brain anatomy, and (iii) generalize to novel stimuli similarly to primates. I will first introduce Brain-Score, our composite benchmark for an extensive comparison of deep nets to primate ventral visual stream. Building on the insights gained by performing such benchmarking, I will describe the CORnet family of models that commits to biological realities of the visual cortex. I will further extend our benchmarking to a much wider image set of images, including cartoons and paintings, to test and compare the limits of generalization in humans and machines. Taken together, our approach brings forward a good baseline deep neural network that could serve as a building block towards developing capable artificial cognitive agents.

Ranking-Based Algorithm for Discrete Facility Location

A. Lančinskas[1], J. Žilinskas[1], P. Fernandez[2], B. Pelegrin[2]

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

The location of facilities is a strategic decision for a firm that competes with other firms to provide goods or services to the customers in a given geographical area. Different location models and solution procedures have been proposed to cope with these problems which vary depending on the ingredients to be considered, such as location space, facility attraction, customer patronizing behavior, demand function, decision variables, etc.

Our research is focused on discrete facility location problems, where all customers in a geographical region are spatially separated into demand points and is already served by some pre-existing facilities owned to different firms. An entering firm wants to compete for this market by choosing locations for new facilities from a finite set of candidate locations.

Despite the well-known binary customers behavior rule in which all customers of a single demand point are served by the most attractive facility, we consider the partially binary rule in which buying power of a single demand point is proportionally divided among all firms choosing the most attractive facility per chain. To ensure viability of the new facilities, the firm includes a constraint for locations: a new facility will be opened only if a minimal market share is captured.

The above facility location problem was solved using the Ranking-based Algorithm for Constrained Discrete Optimization, which is specially adopted for constrained discrete facility location. The algorithm is based on ranking of candidate location depending on their fitness when forming new candidate solutions. The performance of the algorithm was experimentally investigated by solving different instances of the facility location problems using real geographical data of candidate locations.

About Application of Modern Hardware to the Tsunami Wave Parameters Evaluation

M. Lavrentiev[1,2], A. Marchuk[3], A. Romanenko[1,2], M. Shadrin[1]

[1] Institute of Automation and Electrometry, Russia
[2] Novosibirsk State University, Russia
[3] Institute of Computational Mathematics and Math Geophysics, Russia
mmlavrentiev@gmail.com

The up-to-date progress of the IT industry is so amazing that we had single core CPUs yesterday and now we have multi-core chips in out smartphones and even in smart-watches. Now, we can use special accelerators based on FPGA, Intel Xeon Phi and video cards which allows us to have `super computer` under the table. All those achievements allow us to solve more and more complicated problems in real time. In this paper, an overview of the application of modern hardware to the problem of tsunami parameter evaluation is presented.

The only one approach works now to protect human's life and infrastructure along the coast line from a tsunami. It is to inform about the tsunami threat in advance. For this one need to estimate tsunami wave parameters at the epicenter, perform the modeling of tsunami wave propagation and inundation. All those tasks require huge computational power. However, the task gets more complicated since the velocity of the tsunami wave is about 800 km/h in the Pacific Ocean, and if an earthquake occurs near Japan, it takes 20 minutes for the tsunami to reach the nearest point of the dry land. Hopefully, mathematical modeling shows up amazing progress in the recent years due both to a better description of physical processes and availability of the new computer architectures. The authors propose both fast algorithms modeling tsunami wave propagation as well as code acceleration by using advantages of the modern computer architectures.

Scoring Severity of Pulmonary Tuberculosis Based on Automatic Lesion Detection in CT Scans

V. Liauchuk, V. Kovalev, E. Snezhko

Biomedical Image Analysis Department
United Institute of Informatics Problems, Belarus
vitali.liauchuk@gmail.com

Accurate scoring of the severity of pulmonary tuberculosis (TB) is an important problem of quantitative assessment of patients' state. A number of studies were undertaken in order to introduce a scoring system which provide high diagnostic accuracy and reproducibility of TB diagnosis based on chest radiographs and some additional data. With this study we explore the utility of different types of data for predicting cumulative TB severity scores. The data examined include clinical and laboratory data of patients, CT imaging reports provided by radiologists as well as automatically extracted quantitative image features and TB-descriptors derived from detected lung lesions. Automatic detection of lesions in CT scans was performed using state-of-the-art Deep Learning methods and convolutional neural networks (CNN). Image training set for training a CNN-based lesion detector consisted of more than 400 CT scans manually labeled by radiologists. The dataset with known TB severity scores included 279 CT scans. For each patient, a severity score was assigned by a board of medical experts based on a number of factors including TB process prevalence, presence of certain types of lesions, drug resistance status and several others.

The results of experiments on predicting the severity score suggest that automatic detection of lesions in 3D CT scans using Deep Learning methods alone appears to be the most useful. Combination of clinical, laboratory, radiology and conventional image analysis data provides the best achieved accuracy of TB severity prediction. It was proven that distribution and location in lungs of TB lesions automatically detected by Deep Learning is a useful TB image descriptor for a "semantic", medically meaningful, content-based image retrieval tasks. Corresponding algorithms and software are deployed on the dedicated TB web portals of the National Institutes of Health, USA (tbportals.niaid.nih.gov).

This study was supported by the National Institute of Allergy and Infectious Diseases, National Institutes of Health, U.S. Department of Health and Human Services, USA through the CRDF project DAA3-17-63599-1 "Year 6: Belarus TB Database and TB Portals".

Requirements Engineering, Supported by Ontology and Enterprise Modelling

A. Lopata, N. Makrickienė

Institute of Social Sciences and Applied Informatics
Kaunas Faculty, Vilnius University
neringa.makrickiene@gmail.com

In a modern world, as information technology becomes more and more demanding, systems become bigger in the terms of scope, the need in well established requirements becomes crucial. The analysis of an information system requirements should result in the establishment of well-defined functionalities and attributes agreed by the stakeholders. The problem of this research is that, even many tools and methods been already presented in the industry, issues and difficulties still appear in requirements engineering. One of the difficulty is – quality of many specified requirements is poor. In our work, the importance of well prepared requirements is stated by analyzing and merging such technologies as Enterprise Modelling and Ontologies. An Ontology-based, Enterprise Metamodel supported requirements specification tool may help to reduce misunderstanding, missed information, and help to overcome some of the barriers that make successful acquisition of requirements so difficult.

Knowledge-Based UML Models Transformation Algorithm

A. Lopata, I. Veitaitė

Kaunas Faculty, Vilnius University
audrius.lopata@knf.vu.lt

There have been made many efforts for the analysis of UML models generation of diverse knowledge-based models combining workflow patterns, frameworks, modelling languages, and even some natural language specifications.

Researchers interest in UML models in recent years is significantly increased. It is quite challenging to analyse UML models since the knowledge about an enterprise system is spread within separate but related model views. UML models are sustained to decrease the confusion of the problem with the increased number of changes in an enterprise. Knowledge stored in UML models can be effectively implied and can be used in all stages of the information system development life cycle.

UML as one of the major components of IS development life cycle stages models, which can be generated of Enterprise model (EM) as a knowledge repository. This sort of implementation will improve the efficiency of the information system development process participants, such as system analyst, system designer, system developer.

The main scope is to present UML model generation of EM transformation algorithm. The transformation algorithm description is represented in details. Whole generation process steps are defined by particular example following.

Research Action designated as COST Action CA15123: The European research network on types for programming and verification (EUTYPES).

Computer Aided Design of Block Chain System and Effectiveness Analysis of Smart Systems for Information Integration

Z. Mahmood

Department of Software Engineering Informatics
Kaunas University of Technology
zeba.mahmood@ktu.edu

The blockchain is a developing innovation for decentralized and value-based information sharing over an extensive system of untrusted members. It empowers new types of dispersed programming structures, where concession to shared states can be set up without confiding a central integration point. A major trouble for architects designing applications in light of blockchain is that the innovation has numerous arrangements and variations. Since blockchains are at an initial stage, there is little product information or dependable innovation assessment accessible to think about various blockchains. In this paper, we have classified computer aided design of block chain system.

Quantitative Indicators in the Analysis of Functional Styles

J. Mandravickaitė[1,3], T. Krilavičius[2,3]

[1] Vilnius University
[2] Vytautas Magnus University
[3] Baltic Institute of Advanced Technology
justina@bpti.lt

Text analysis and information retrieval focused on topic of a text have been common in research works. However, a text also contains useful information in its style. Textual style can be described as a pragmatic meaning, referring to such aspects as cohesion, function, interpersonal distance, etc. Different choices within these meaning systems define stylistic variation. This view corresponds to works in computational stylistics, e.g., authorship attribution, and suggests new research areas, e.g., automatic detection of rhetorical purpose. We analyzed texts from functional style perspective, considering their communicative aspects. Discriminant analysis with chosen features enabled characterization of stylistic differences between groups of texts in terms of functional style that can be linked back to different textual functions.

We applied 4 quantitative indicators (Activity, MATTR, Lambda and Thematic concentration) for the analysis of texts of different functional styles. Functional style is a variety of standard language which can be defined via 5 characteristics: field usage, content, text functions, stylistic devices and linguistic means. For our experiments we used Lithuanian corpus MATAS comprised of 1.6 million words. We analyzed texts of 4 functional styles: publicistic (P), administrative (A), scientific (S), belles-lettres (B).

Indicators used in this study, even though do not include all features of functional styles, allowed to observe certain tendencies. MATTR, Lambda and Thematic concentration were statistically significant at least in some cases, Activity was insignificant in all cases. Based on conformity indicator C, texts of B and P styles conformed to their style tendencies the most while texts of A style were the most "original". Finally, mixture discriminant analysis with 3 statistically significant indicators classified 75 texts out of 92 correctly (P: 29/30, A: 34/40, S: 9/13, B: 3/9).

Investigation of User Vulnerability in Social Networking Site

D. Mažeika, J. Mikejan

Vilnius Gediminas Technical University
dalius.mazeika@vgtu.lt

Vulnerability of social network user becomes a social networking problem. A single vulnerable user might place all friends at risk therefore, it is important to know how security of the user can be improved. In this research, we aim to address some issues related to user vulnerability to phishing attack. Text messages of the social network site users were gathered, cleaned and analyzed. Moreover phishing messages were build using social engineering methods and sent to the users. K-means and Mini Bach K-means clustering algorithm were implemented to make text message mining and to find the keyword used for phishing message. Moreover, special tool was developed to automate this process. Analysis of users responses to the phishing messages formatted applying different clustering algorithms and social engineering methods was performed and corresponding conclusions about user vulnerability were made.

Fuzzy Balancing in the Planning of the Quality of Enterprise Business Services

J. Miliauskaitė, L. Paliulionienė

Institute of Data Science and Digital Technologies
Vilnius University
jolanta.miliauskaite@mii.vu.lt

The modelling of the quality of service (QoS) is one of the important issues in service-oriented enterprise system (SoES). QoS is a complex and multi-sided concept, and a lot of different stakeholders often differently understand the SoES service quality. A view-based framework that uses different viewpoints and perspectives for the evaluation of the quality of enterprise business service (QoS_{EBS}) was proposed in [1]. According to this framework, the QoS_{EBS} planning problem can be decomposed into subproblems: fuzzification, fuzzy balancing, fuzzy reasoning, perspective aggregation, views aggregation, and linguistic approximation.

This work presents a problem-independent methodology for fuzzy balancing and demonstrates the applying of this methodology for QoS planning.

[1] Lupeikienė, A., Miliauskaitė, J., Čaplinskas, A. A Model of View-Based Enterprise Business Service Quality Evaluation Framework. Informatica, 24(4), 2013, p. 543–560.

On the Implementation of Three-Stage Algorithm for EEG Classification by Diagnosis

A. V. Misiukas Misiūnas[1], T. Meškauskas[1], R. Samaitienė[2,3]

[1] Institute of Computer Science, Vilnius University
[2] Children's Hospital, Affiliate of Vilnius University Hospital Santaros Klinikos
[3] Clinic of Children's Diseases, Faculty of Medicine, Vilnius University
andrius.misiukas@mif.vu.lt

A three step algorithm for electroencephalogram (EEG) classification by diagnosis and its implementation details is discussed. 94 EEGs of children (3-17 years old) are under classification between two different epilepsy types. All EEGs with known diagnoses are provided by Children's Hospital, Affiliate of Vilnius University Hospital Santaros Klinikos. Patients from Group I are diagnosed with rolandic epilepsy, patients from Group II – with structural focal epilepsy. The algorithm consists of three main steps: 1) EEG spike detection, 2) evaluation of spike parameters, 3) classification of EEGs (between Group I and Group II) by machine learning (ML) methods. Spike detection was implemented in both multi- and single-threaded computing environments. EEGs are classified by evaluated parameters of detected spikes. Since artificial neural network (ANN), as well as most other classification algorithms, require a fixed amount of inputs, EEGs are divided into fragments containing 100 spikes. This also helps to overcome problems related to having relatively few EEGs. The algorithm is implemented in Python 3.6 programming language and Scikit-learn ML library as well as NumPy and SciPy. Different ML methodologies (in the third step of the algorithm), were compared: a) Decision tree, b) Random forest, c) Extremely randomized tree, d) Adaptive boosting (AdaBoost), e) Supported vector machine (SVM), f) Linear discriminant analysis (LDA), g) Logistic regression, h) Naïve Bayes, and i) ANN. Analyzing various ML quality metrics, some methodologies found to be guessing the answer by making benefit of uneven distribution of patients between Group I and Group II. ANN based classifier prevailed as best suited methodology according to the true positive rate, true negative rate, and other metrics, achieving approximately 73% accuracy.

Accelerating the Dose Evaluation for Intensity Modulated Radiotherapy

J.J. Moreno[1], J. Miroforidis[2], E. Filatovas[3], I. Kaliszewski[2], E. M. Garzón[1]

[1] Informatics Department, University of Almería, Spain
[2] Systems Research Institute, Polish Academy of Sciences
[3] Institute of Data Science and Digital Technologies, Vilnius University
juanjomoreno@ual.es

The planning of Intensity Modulated Radiotherapy (IMRT) is an effective cancer treatment if the parameters related to the planning are optima. To design effective IMRT planning, it is necessary to evaluate the radiation doses in an efficient way. There are several dose models, most of them based on the computation of products of deposition matrices (defined for every geometric configuration of every beam) by the vector x whose elements define the radiation time of every beamlet. The deposition matrices are sparse because every element represents the contribution of one particular beamlet (column-index) to the radiation dose for every voxel (row-index). This way the deposition matrices are stored with compact formats to avoid storage and computation of zero elements. The radiation doses of every voxel are computed by the addition of several sparse matrices products. This procedure consumes large computational resources since for fine spatial discretization meshes, the tomographic data and deposition matrices require very large capacity of memory and the number of Floating Point Operations related to the corresponding computation is also very high. The goal of our work is the acceleration of these operations by: (1) the exploration of several schemes to define the matrix operations involved in the computation of the radiation doses of every voxel for one or several particular IMRT plans and (2) the exploitation of modern High Performance Platforms such as multicore and GPUs.

Deep Learning-Based Method for Quantitative Collagen Framework Analysis in Routine Pathology Images

M. Morkūnas[1,2], P. Treigys[1], A. Laurinavičius[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] National Center of Pathology
Affiliate of Vilnius University Hospital Santaros klinikos
mindaugas.morkunas@mii.vu.lt

Collagen fibers are the most primitive structural elements of connective tissue. The fine structure of this tissue changes within many human pathological conditions and especially solid tumor cancers. Therefore, collagen detection in histological sections and its detailed quantitative characterization is of great importance for the classification of different pathological states. Many methods used to visualize collagen structure in a tissue specimen such as second harmonic generation microscopy, electron microscopy, atomic force microscopy are limited by the relatively small field of view, complicated sample preparation, and often a high cost of equipment. In this context, the cheaper bright-field microscopy methods seem to be neglected due to the fact that fibrous collagen structure cannot be extracted with the amount of precision offered by specialized imaging techniques.

In this study, we used a deep learning neural network to train on a minimal amount of collagen annotated routine bright-field microscopy images of breast carcinoma. Using the trained model we extracted important collagen framework features from Sirius Red-stained images of 93 breast carcinoma tissue microarray cores. Extracted features significantly relate to the clinical outcome of patients. We also effectively applied the same model to a variety of different tissues, as well as staining protocols without further retraining.

Smart System Prototype of Optical Resistance Testing for Laser Technology Sector

D. Navakauskas[1], R. Martavičius[1], T. Sledevič[1], J. Skirelis[1], A. Melninkaitis[2], M. Ščiuka[2]

[1] Vilnius Gediminas Technical University
[2] JSC "Lidaris"
julius.skirelis@vgtu.lt

Optical elements of industrial lasers degrade due self-generated intense light. Accordingly, optics manufacturing technologies such as glass polishing or deposition of optical coatings should be optimized for high laser damage resistance. Such optimization, based on trial and error approach, demands thousands of laser damage tests. To make laser damage testing widely accessible existing test benches should be completely re-engineered to enable hands free operation. Appropriate damage detection and inspection systems should be developed and integrated into all motorized measurement head.

The goal of the project was to develop a next generation prototype of laser damage testing metrology system with built in all the required key elements (including reliable video based damage detection, Nomarski based microscopic inspection, online beam characterization system and all-motorized positioning system) enabling fully automated hands-free operation of damage testing process. The main results of the project are:

- Redesign of all optomechanical parts required for complete automation of sample positioning and exposure by laser thus eliminating human motion waste operations.
- Development of computer vision based laser induced damage detection unit that is capable to stream and process video signals at least with 100 Hz repetition rate.
- Integration of on-demand Nomarki (DIC) microscopy based video inspection unit into measurement head.
- Development of autofocusing algorithm for microscopic damage inspection.

Within the project multiple CMOS high resolution image sensors were integrated. The prototype was implemented using real-time systems and FPGA programming techniques. Laser beam diameter estimation, laser beam tracking and laser induced damage detection FPGA IP cores were developed.

Research and development was funded by Agency for Science, Innovation and Technology through technology development project No. TPP-01-011.

Human-Computer Interaction: A Behavioural Approach

P. Novais

Department of Informatics
School of Engineering
University of Minho, Portugal
pjon@di.uminho.pt

Academia and enterprises have been very prolific in researching human-computer interaction and the human behaviour issues through controlled studies, surveys, laboratory prototypes and applied case-studies. The results of this research efforts show that there are several key factors involved in Human-computer Interaction activities. These include physiological and psychological indicators, as well as behavioral and even somatic ones.

Stress is a critical component of the human activity, and it shows through physical, mental, or emotional tension, having a high impact in medical or biological contexts. Thus, stress has a great repercussion in psychological conditions such as depression and anxiety. In a corporate environment, workers are under increasing demand for performance, which leads to being under constant pressure. Therefore, their level of stress is greatly affected, presenting significant variations. To tackle these issues, this research aims to produce a new approach in human–computer interaction by developing a distributed multi-modal framework to monitor and assess the psychological stress of people doing high-end computer-related tasks. This is done in a non-intrusive and non-invasive way, using soft sensors to monitoring activities (e.g., task performance and human behaviour).

An Efficient Software for Hyperspectral Classification Using Isometric Mapping

F. Orts[1], G. Ortega[2], E. Filatovas[3], O. Kurasova[3], I. García[1], E. M. Garzón[1]

[1] University of Almeria, Spain
[2] University of Málaga, Spain
[3] Institute of Data Science and Digital Technologies, Vilnius University
francisco.orts@ual.es

Hyperspectral images (HSIs) collect information about the spectral signatures of the materials, with hundred or even thousands of bands covering specific lengths of wave. They have been used in many applications of remote sensing and beyond. However, the high volume of information to process makes necessary to discard irrelevant information to analyse data in an efficient way. In this context, Isometric mapping (Isomap) algorithm is one of the earliest approaches to manifold learning and it is often when analysing hyperspectral images. Isomap allows to reduce such hyperspectral images from a high-dimensional space into a lower-dimensional space, keeping the critical original information. It is a nonlinear generalization of the Multidimensional Scaling (MDS), where the main idea is to perform MDS in the geodesic space of the nonlinear data manifold. There are several approaches to solve MDS problem.

This paper studies one of the most accurate MDS, the SMACOF algorithm. Using well-known hyperspectral images, a deep comparison in terms of accuracy is carried out between Isomap implementations based on SMACOF and Classical Scaling MDS. A support vector machine for hyperspectral images classification has been used to test the obtained results in both cases, showing than better accuracy has been reached with Isomap based on SMACOF. Finally, due to the high computational cost of SMACOF MDS, the use of High Performance Computing is mandatory. Then, a parallel version of Isomap based on SMACOF MDS has been implemented using GPU computing and evaluated using a set of HIS tests.

Aspect-Based Sentiment Analysis with Word Embeddings for Lithuanian

M. Petkevičius, D. Vitkutė-Adžgauskienė, D. Amilevičius

Vytautas Magnus University
daiva.vitkute-adzgauskiene@vdu.lt

Opinion mining (sentiment analysis) problem is usually solved by applying a lexicon-based model (e.g., in www.semantika.lt project). Deep neural networks are gaining popularity in different text processing and classification tasks, sentiment analysis being one of them. The paper presents research on deep-learning algorithms for aspect-based sentiment analysis, aiming at sentiment definition for each identified aspect of a given entity (e.g., a person, a brand name, etc.). Most of the distributional word embedding models nowadays learn semantic representations of words ignoring their morphological structure. This appears to be a limitation, especially for languages with complex morphology, as their vocabularies are quite large, most words are infrequent, resulting in models being unable to learn acceptable semantic representations. The paper presents the research results on the application of a FastText n-gram (subword) based word embedding model, treating each word as a composition of character n-grams (subwords). Subword-level information is crucial for capturing word meaning and morphology, especially for out-of-vocabulary (OOV) entries and rare words. Comparison with the traditional word2vec (Skip-Gram) word embedding model showed, that the subword feature enhances learning for morphologically rich, heavily inflected languages, such as the Lithuanian language. It is shown that for Lithuanian texts FastText models with n-grams do significantly better on syntactic tasks (aspect-sentiment pair), because of the syntactic questions being related to morphology of the words. However, original word2vec models perform better on semantic tasks, since words in semantic analogues (distributional hypothesis) are unrelated to their n-grams, and the add-on information from irrelevant character n-grams can worsen the embeddings. Despite having more OOV words to predict, FastText model showed better results over traditional Skip-Gram model. The recommended aspect-based sentiment analysis approach exploits joint effect of semantic and syntactic word embeddings by taking advantage of semantic word embeddings and syntactic word embeddings at the same time.

Finite Elements Modeling of Electrochemical Impedance Spectra of Defected Phospholipid Membranes

T. Raila[1], T. Meškauskas[1], G. Valinčius[2], T. Penkauskas[2], M. Jankunec[2]

[1] Institute of Computer Science, Faculty of Mathematics and Informatics
Vilnius University
[2] Life Sciences Center
Vilnius University
tomas.raila@mif.vu.lt

Tethered bilayer lipid membranes (tBLM) are popular experimental platforms for studying protein-membrane interactions. One of alternating current (AC) techniques used to assess dielectric properties of such membranes is electrochemical impedance spectroscopy (EIS). While this method is useful for determining macroscopic properties of bilayers, it provides no direct information on structural properties of membranes containing defects. Such cases often require more complex microscopy techniques, such as atomic force microscopy (AFM).

The goal of this study was to investigate the relation between EIS spectral features and structural properties of defected membranes. We applied finite element analysis (FEA) technique to model EIS spectra for various defect distributions. Three-dimensional membrane models were implemented and solved with COMSOL Multiphysics FEA software. Calculations were carried out in different computing environments and dependency between model complexity and calculation times was investigated.

Both experimentally registered by AFM and computer-generated defect distributions of varying density and defect size were used in modeling. Comparison of modeled EIS spectra for heterogeneous defect distributions and analytical EIS solutions of homogeneous distributions indicated only quantitative differences between the two approaches. EIS spectra modeled with experimentally registered defects and computer-generated distributions of equivalent density exhibited discrepancies in cases of defect clusters present in experimental data. Comparing experimentally measured EIS spectra of real membranes with modeled spectra revealed significant correlation between EIS spectral features of both datasets.

The Effect of Student-Oriented Teaching Practices in Explaining PISA Science Performance Across EU Countries

S. Raižienė, D. Stumbrienė, L. Ringienė,
R. Dukynaitė, A. Jakaitienė

Vilnius University
saule.raiziene@fsf.vu.lt

The acquisition of scientific principals and theories during school years increasingly gains greater value as science-related employment is expected to grow (Fayer, Lacey & Watson, 2017). However, the students' interest in science is declining (Potvin, Hasni, 2014). Therefore it is important to analyze what student-oriented teaching practices are the most effective for science disciplines in order to keep students' interest in science. Teaching practices vary across educational systems and their effectiveness depends on the composition of other factors (Kyriakides, 2008). This study investigates how differ and what is the differential effectiveness of student-oriented teaching practices (enquiry-based science teaching practices, adaption of instruction, teacher support and feedback) across EU countries. Data from PISA 2015 were used. 24 EU education systems were analyzed. Multiple linear regression analyses for each education system was performed with student's science performance as the dependent variable and four students' perceived student-oriented teaching practices (enquiry-based teaching, adaption of instruction, teacher support and feedback) as independent variables. Student economic, social and cultural status was entered as control variable. From the study we conclude that the prevalence (based on students' reports) of four analyzed student-oriented teaching practices is different across EU countries. Similar patterns of association for adaption of instruction and perceived feedback with science performance, and different patterns of association for teacher support and enquiry-based teaching with science performance are observed across EU countries. This confirms different effectiveness of student-oriented teaching practices in EU learning contexts.

Hierarchy Based Competences' Analysis: Opportunities for Personalized E-Evaluation

S. Ramanauskaitė, A. Slotkienė

Vilnius Gediminas Technical University
asta.slotkiene@vgtu.lt

The growth of information availability changes the modern concept of learning. Previously it was expected from the teacher to provide the main source of learning content. However, currently the teacher is gradually replaced or supplemented by e-learning systems. In order to ensure objective evaluation of students competences, the study process results must be discrete and semantically expressed. However the current system, when student competences are summarized and expressed as one quantitative metric (mark), does not express the list of students competences and their level. In order to solve the problem in this paper we propose a method for the design of the competence tree. The competence tree has to be formatted based of context modelling principles and analysis of Scope-Commonality-Variability. The usage of competence tree for student competence evaluation proposes a clearly defined and semantically expressed evaluation method for both - human and e-learning evaluation process. Results of the empirical experiment of adapting the proposed competence tree design and application for competence e-evaluation method, based on flexibility, adaptability and granularity of learning material are presented in this as well.

Distributed of Genetic Algorithm for Optimization of Piles

M. Ramanauskas, D. Šešok

Vilnius Gediminas Technical University
mikalojusrama@gmail.com

The grillage foundation is a popular foundation type, especially in the case of weekly ground. Reinforced concrete piles are the terminal element of the erection, which distributes the loadings coming from the erection through the connecting beams. The determination of the minimal number of piles and their placement at the right positions is a relevant engineering problem.

Despite the significant advance in optimization methods and computer hardware, optimization of piles still requires unacceptably long computing time due to the large number of required numerical experiments. Naturally, to solve the problem of global optimization, parallel and distributed computing has become an obvious option for significantly increasing computational capabilities.

The voluntary BOINC (Berkeley Open Infrastructure for Network Computing) projects present a cheap alternative, which can provide computational resources sufficient for optimization of grillages. Currently, it is the most popular open source middleware for deploying distributed computing applications on volunteered computing devices connected over the Internet.

The paper provides a distributed computation method for genetic algorithms. The method was realized on the BOINC platform and the analysis of the results.

On the Visual Approach to Global Optimization for Multidimensional Scaling

M. Sabaliauskas, G. Dzemyda

Institute of Data Science and Digital Technologies
Vilnius University
martynas.sabaliauskas@mii.vu.lt

We are presenting a new geometric-based method for a visual representation of high-dimensional data. The solving problem is to visualize a given multidimensional data to lower-dimensional space by holding similarity and distances between multifaceted objects as much as possible. Our proposed optimization method consists of the local and the global optimization phases. The local optimization phase uses a center of a special gradient-circle as a seed to start a local geometrical optimization for finding an optimal position of the unfixed lower-dimensional vertex. The global optimization phase respectively corrects positions of all optimizing lower-dimensional vertices until a value of the global stress function stops decreasing. According to experimental results, our method tends to solve the global optimization problem of multidimensional scaling.

Multicriteria Decision Making Model for Anomaly Detection in Financial Markets

V. Sakalauskas, M. Liutvinavicius, D. Kriksciuniene

Vilnius University
dalia.kriksciuniene@knf.vu.lt

Multicriteria decision making model is proposed for anomaly detection in financial markets. The factors characterizing financial and behavioral patterns are explored and integrated to the dynamic dashboard simulation. The proposed algorithms enable to visualize market status, alert of the approaching anomalous situations and make investment decisions outperforming the state of the art models.

Blockchain Technology for Security and Privacy in Internet of Things

R. Savukynas, V. Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University
raimundas.savukynas@mii.vu.lt

The Internet of Things (IoT) is a global network of objects with sensors, controllers, software, and capable of gathering, processing, exchanging information, which functioning is based on interactions between existing or completely newly developed information and communication technologies. The global network allows linking various types of objects over the internet, establish their operation rules, protect outgoing confidential information, ensure integrity of received data, and lower consumption of energy resources. Although the provided IoT benefits are unlimited, there are many security and privacy challenges facing adopting the IoT in the real world due to limitations of classical client and server model. This model requires all objects to be connected through the centralized server, which creates a single point of failure, so moving the IoT into the decentralization system is the right decision. One of the popular decentralization systems could be based on blockchain a powerful technology that decentralizes computation and management processes, which can solve security and privacy issues in IoT. The blockchain technology has a great potential in the most diverse technological areas and can significantly help achieve the IoT view in different aspects, increasing the decentralization, enabling transaction models, and allowing autonomous coordination of the objects. The decentralization system of the blockchain has the ability processing of billions of transactions between IoT objects, significantly reduce costs associated with installing and maintaining large centralized data centers, distributed computation and storage needs of millions of objects that compose IoT networks. The blockchain technology eliminates the single point of failure associated with the centralized server and integrating blockchain with IoT will allow the peer-to-peer messaging, file distribution and autonomous coordination between objects with no need for the centralized client and server model. In this work, we investigate the concepts of IoT security and privacy, a solution based on blockchain technology is proposed.

Strategy Selection for Autonomous Robot Area Exploration by Multi-Criteria Decision Making Approach

R. Semėnas

Department of Graphical Systems
Vilnius Gediminas Technical University
rokas.semenas@vgtu.lt

The autonomous robot navigation problem has been a prominent study subject in the past ten years. Various methods for the unknown area exploration were proposed, however, the aspect of robots' safety is rarely addressed. In the present study, different autonomous robot safe area exploration strategies are compared by multi-criteria decision making approach. In the context of this work, the studied strategies correspond to separate alternatives that are evaluated to the relation of the proposed criteria list. The created autonomous robot system is experimentally evaluated.

Application of Machine Learning to Analysis of M. Tuberculosis Whole Genomes and Investigation of Drug-Resistance Profile

R. S. Sergeev[1], A. V. Tuzikov[1], A. Gabrielian[2], A. Rosenthal[2]

[1] United Institute of Informatics Problems NASB
Minsk, Belarus
[2] Office of Cyber Infrastructure and Computational Biology
NIAID, Bethesda, USA
roma.sergeev@gmail.com

Emergence of drug-resistant microorganisms has been recognized as a serious threat to public health worldwide. This problem is particularly discussed in the context of tuberculosis treatment.

Alterations in pathogen genomes are among the main mechanisms by which microorganisms exhibit drug resistance. Despite the diversity of the mutations investigated so far, they do not fully explain all cases of drug resistance observed. For example, analysis of TBDreamDB database and GenoType MTBDRplus/MTBDRsl assays has shown that only 85.7% of ofloxacin and 51.9% of pyrazinamide resistance could be explained by the presented high-confidence mutations. For other second-line and add-on drugs (ethionamide, kanamycin, amikacin, capreomycin, cycloserine) these values are even lower. This indicates that the genetic basis of drug resistance is more complex than previously anticipated.

Modern machine learning methods can provide a powerful tool for making sense of massively generated genomic data. Despite there are still many challenges on this way, comprehensive analysis of multi- and extensively drug-resistant tuberculosis may become very valuable for understanding patterns of parasite evolution under treatment and host immune pressure.

Within this study our efforts were focused on:

1. Development of bioinformatics pipelines for M.tuberculosis whole-genome assembly and annotation.
2. Comparative analysis of pathogen genomes to investigate features that may explain virulence and contribute to drug resistance.

3. Applying machine learning methods trying to explain missing heritability of resistant phenotypes.

As a use case, we analyzed whole genomes of 738 M.tuberculosis strains isolated in Azerbaijan, Belarus, Georgia, Moldova and Romania. The dataset comprised drug-sensitive, multi drug-resistant, pre-extensive drug-resistant, extensive drug-resistant and totally drug-resistant strains sequenced on Illumina platform.

Our work has identified important points applicable for analysis of drug-resistant pathogens. All genomes were submitted to Genbank and linked to anonymized patients' data (clinical, socioeconomic, SNPs, chest X-Ray and CT) available on M/XDR-TB Portals network via https://tbportals.niaid.nih.gov/.

Maximal Zerocross Density Decomposition: A Novel Signal Decomposition Method

T. Sidekerskienė[1], R. Damaševičius[2], M. Wozniak[3]

[1] Department of Applied Mathematics
Kaunas University of Technology
[2] Department of Software Engineering
Kaunas University of Technology
[3] Institute of Mathematics
Silesian University of Technology, Poland
tatjana.sidekerskiene@ktu.lt

We developed the Maximal Zerocross Density Decomposition (MZDD) method for decomposition of signals into intrinsic modes. We discuss the main properties of MZDD and parameters of MZDD modes (statistical characteristics, principal frequencies and energy distribution). A modal analysis of a nonstationary signal is provided.

Peculiarities of the Thermal Expression of Students' Feeds, Linking Them to the Consequences of Sport Traumatic Events

D. Skaudickas[1], A. Skaudickas[2], V. Veikutis[1]

[1] Lithuanian University of Health Sciences
[2] Gymnasium of the Lithuanian University of Health Sciences
darijusskaudickas@gmail.com

The onset of common traumatic disorders is usually difficult to detect. Pain during or after physical activity is usually associated with natural status and is mostly compensated by using of anti-inflammatory drugs. Diagnostic investigations are usually not performed for economic, social or other reasons, what developed disease continues and lead to orthopedic pathology progress especially for adolescents. Thermography could be informative method for more detailed diagnostics and treatment efficacy assessment. The aim of the study was to determine the peculiarities of temperature distribution of students foots and linking them to existing or past sports traumas.

52 adolescents were included into study (permission BC-MF-205; LUHS Bioethics Center). They were 15-16 years old, of them boys were 61.5%, (n = 32), and girls – 38.5%, (n = 20). The thermographic scans were performed by a thermocamera FLIR T440, images were processed using FLIR Tools 2.0 analyzing system.

The main both foot temperature range detected in martial arts students, reflecting the great traumatic status especially in small joints and fingers. The higher the right leg (support) the trend of failure prevails with basketball and volleyball players – they are referred to the permanent right leg pain. Dramatically broke up found in martial arts and football players temperature variance. Representatives of these sports suffer traumas in all of the areas of the feet – there dominated combustion of the foot, inability to move, fatigue symptoms. Ankle injury significantly broke up with volleyball and basketball sport subjects and the most vulnerable are not the backing (usually the right), but the left

leg. Finding the temperature of the inflammation in individuals fully reflects the objectively expressed pain or discomfort of movement.

We don't find statistically significant temperature differences between sports and non-sports subjects. The temperature resolution of inflammation to individuals fully reflect the objective expressed pain by the movement or complaints of discomfort. The most traumatic are basketball (particular in the left foot), football and martial arts (both legs) sports. The objective of the survey complaints well correlated with thermography findings.

Edge-Based, Dynamic, Hierarchical Architecture for Super-Resolution Local Climate Prediction Using a Swarm of Field Programmable Gate Arrays as Accelerators

L. Stasytis, A. Uselis, I. Gaubas, K. Bagdonas

Kaunas University of Technology
lukas.stasytis@ktu.lt

We present a novel, Edge-based, dynamic, hierarchical architecture using a swarm of neural networks (NN's), aimed at the implementation of super-resolution algorithms for processing climate data. The architecture features Field Programmable Gate Arrays (FPGA's) organized in a hierarchical structure to act as accelerators by providing hardware level data processing speeds while preserving on-the-fly reconfiguration of the software models using software-to-hardware converters. The current capacity of FPGAs acting as nodes in a swarm to predict future climate using dynamically changing input data is explored. Input data includes the European Earth Observation Copernicus program datasets and APIs, edge device readings and other, local, FPGA node predictions with temperature and humidity being the key input and output data points.

Improved DIRECT-Type Algorithm for Constrained Global Optimization Problems

L. Stripinis, R. Paulavičius

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

In this talk, we consider a general global optimization problem. The well-known derivative-free global-search DIRECT (DIvide a hyper-RECTangle) algorithm performs well on a subclass box-constrained problems. Unfortunately, quite often the efficiency of the DIRECT deteriorates on problems with many local optima or when the solution with high accuracy is required. To overcome these difficulties, different regimes of the global and local search are introduced, or the algorithm is combined with a local optimization scheme. We investigate a different direction of addressing DIRECT inefficiencies and propose a new strategy for the selection of potentially optimal rectangles. Our novel approach does not require any additional parameters or local search subroutines.

Solving problems with general constraints, DIRECT does not naturally address them, and till now only a few DIRECT extensions were proposed. Thus, in this talk, several different general constraints handling methods within the DIRECT framework are proposed and compared with recently proposed alternative extensions. An extensive experimental investigation reveals the effectiveness of the introduced enhancements.

Western Union: *Moving Money for Better* with Robotic Process Automation (RPA)

Š. Šuipis

VP, Operations & Managing Director
Western Union Processing Lithuania

- Lessons learned from the global roll out of RPA at Western Union.
- Identifying key tasks and processes to be automated through RPA.
- Developing and adjusting a global RPA roadmap.
- Linking RPA with Business Process & Human Resources Management for an integrated automation framework.

Cloudlet Based Prediction of Energy Consumption for Mobile Phones

J. Toldinas, B. Lozinskis

Faculty of Informatics
Kaunas University of Technology
eugenijus.toldinas@ktu.lt

The number of mobile phone users in the world is expected to pass the five billion mark by 2019. Today mobile phone in our pocket is a high performance computer with gigabytes of RAM, processor that has two, four or more cores, Wi-Fi, Bluetooth, display, and etc. All of mentioned devices consumes battery energy. Unfortunately, while technical parameters of mobile phones growth quickly, battery lifetime remains very short and users must charge their phones more and more frequently. BYOD (bring your own device) is the increasing trend when employee brings their owned smartphones, tablets or laptops to their workplace. That is why prediction of energy consumption for devices with battery is an important issue.

We propose a novel cloudlet based prediction method of energy consumption for mobile phones. The main contributions of proposed method are twofold. First, we propose a smart model for combining mobile phone usage and energy consumption statistical data. Second, the cloudlet-based system uses predictive modelling and statistical data of our proposed model to make predictions and inform mobile phone user about future energy consumption.

Dynamical Network Modelling Based on Triadic Closure Effect

R. Užupytė[1,2], E. C. Wit[3]

[1] Vilnius University
[2] Baltic Institute of Advanced Technology
[3] University of Lugano, Switzerland
ruta.uzupyte@bpti.lt

Statistical network analysis has become an important methodological framework for modeling and analysis of complex relational data. A huge diversity of complex systems has a non-static structure of connections as the links between objects dynamically change over time. Many scientists believe that triadic closure effect is a fundamental mechanism of dynamic network formation and evolution. The hypothesis of triadic closure contends that new links in a network arise preferentially between nodes having common neighbors. This study focuses on the possibility to incorporate connectivity data into the framework for temporal network analysis. The proposed approach is based on stochastic actor-based modelling framework introduced by Snijders [1, 2]. Proposed model describes changes in network structure as an actor-oriented Markov process. Moreover, it has been shown [1] that parameters in such a model can be estimated using the method of moments.

[1] Snijders, T. A. B. Stochastic actor-oriented models for network change, Journal of Mathematical Sociology, 21(12), 1996, p. 149-172.

[2] Snijders, T. A. B. Models for longitudinal network data, Models and Methods in Social Network Analysis (Cambridge University Press, New York), 2005, p. 215–247.

Effectiveness and Efficiency of Education Systems: A Review of Research and a Way Forward

J. Vaitiekaitis, D. Stumbrienė, A. Jakaitienė

Vilnius University
dovile.stumbriene@mii.vu.lt

The aim of our research (EFECTAS: Effectiveness and Efficiency Analysis of Education Systems in EU Countries Employing Secondary Big Data) is to investigate the factors determining the effectiveness and efficiency of the EU education systems. The presentation will focus on the discussion about differences between the effectiveness and efficiency in education, practical examples of variables and models actually used in empirical studies. We will discuss distinct theoretical approaches between effectiveness (the levels of educational outputs achieved) and efficiency (the levels of educational inputs vs outputs) in education, and provide an overview of input and output variables, as well as different targeted education system levels and its classifications. Finally, we will present traditional frontier approaches (Data Envelopment Analysis) for analysing effectiveness and efficiency in a cross-country perspective.

A Deep Knowledge Based Evaluation of Applications Interoperability

A. Valatavičius, S. Gudas

Institute of Data Science and Digital Technologies
Vilnius University
valatavicius.andrius@gmail.com

Interoperability of the enterprise applications in the dynamic environment is a complex issue. New methodological approaches and solutions are required. The background of our approach is the casual modeling paradigm integrated with MDA approach, combining the business process modeling and control theory principles, enterprise architecture modeling, and autonomic computing principles. The enterprise application software (EAS) interoperability capability evaluation method is presented. Management transaction concept reveals a causal dependencies and the goal-driven information transformations of the enterprise management activities (a deep knowledge). An assumption is that autonomic interoperability is achievable by gathering knowledge from different sources in an organization particularly enterprise architecture and software architecture analysis through web services can help gather required knowledge for automated solutions. The prototype version for testing of enterprise application integration solutions is under development. The background of the interoperability capability evaluation method is a deeper look into evaluation potentiality of interoperability by comparing edit distance of web service operations, objects anf fields gathered for each enterpise application software. The interoperability of few enterprise application software systems (SuiteCRM, ExactOnline, NMBRS, Prestashop) have been indicated by comparing webservice operations using edit distance calculations. The edit distances have been calculated to gather data for for evaluation potentiality of the interoperability solution.

Aspects of Data Collection for Abnormal Marine Transport Evaluation

J. Venskus[1,2], P. Treigys[1], J. Bernatavičienė[1], A. Andziulis[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Klaipėda University
julius.venskus@mii.vu.lt

Highly-loaded seaports have extremely complex and intensive marine vessel traffic, which generates large volumes of traffic data. Meteorological conditions influence maritime traffic, so they must also be taken into account in order to train the model capable of recognizing the abnormal movement of the sea transport. Typically, meteorological data provided by a number of data providers differs in a given region, so data must be processed beforehand to achieve better model accuracy. This paper reviews methods of obtaining vessel traffic and meteorological data in regions such as Rotterdam, Gdansk, Malmö, Copenhagen, Kiel and Lübeck. Also, it discusses traffic data structure and types, data conversion to Euclidean space and data filtering. Author describes possibilities to obtain technical parameters of ships and existing limitations while taking into account the available vessel data. By assessing the characteristics of the received traffic data, the regionalization of the meteorological data and the assignment to a specific maritime traffic data vector by using the method of the closest neighbor, depending on the time of traffic in a given region, is presented. Also, the statistical indicators of traffic data are presented. Traffic intensity is evaluated and visualized by applying the Gaussian kernel density estimation function.

Data Analysis in Setting Action Plans of Telecom Operators

M. Visan[1], A. Ionita[2], F. G. Filip[3]

[1] Department of Engineering, Mechanics, Computers
Romanian Academy, School of Advanced Studies
of the Romanian Academy (SCOSAAR)
[2] Romanian Academy Research Institute for Artificial Intelligence
"Mihai Drăgănescu" (ICIA)
[3] Romanian Academy
maria.visan@ingr.ro

At present we can notice that there is a fierce battle of telecom operators to win the communication service market. The telecom market is currently in full development and opens opportunities for both operators, institutions or the public potentially consuming such services. Telecom operators are particularly well positioned to begin to capitalize on this opportunity, their relationship with customers, assets, and build a distinct position in the market. Thanks to market novelty, this is the optimal time for telecom operators for defining suitable value proposals, structuring the right technologies and "go-to-market" partnerships. Fortunately, they possess huge volume of data which can analysed with a view to serve for preparing good decisions.

For a more in-depth assessment, this paper proposes a detailed analysis of this "battlefield", inlighting the context, the issues of telecom operators that support data for these services, examples of services and the potential users of these services, possible services solutions architectures and methodology implementation. All these aspects will be exemplified by using practical results.

Non-Standard Distance Measures in Data Streams Classification

K. Ząbkiewicz

Department of Economics and Informatics
University of Bialystok, Poland
kamil.zabkiewicz@gmail.com

Data streams are recently getting more and more attention. We can see their presence in marketing (e.g., analysis of customer's buying activity), healthcare (e.g., measuring heartbeat by the wearable sensors to predict heart attack), behaviour analysis (e.g., sentiment analysis of the messages in Twitter) or computer network security (e.g., analysis of the network packets to detect the intrusion). Classification of this sort of data is a very challenging problem.

The first measure we are dealing in this work is the Normalized Compression Distance (NCD). The main reason for choosing this method was that the NCD is parameter-free, feature-free, and alignment-free. There is no need to set various parameters, such as learning speed, the number of epochs, weights for features etc.

The second non-standard distance measure considered in this work is called Lempel-Ziv Jaccard Distance (LZJD). It is also parameter free and works with the raw binary data. What is more, this method is the real distance measure, satisfying triangle inequality and is computed faster than NCD. The basis of this approach is the computation of the minimum hash that leads us to Jaccard similarity measure. This similarity can be easily converted into the distance by subtracting it from 1.

The primary motivation of this work is to test how these measures work in data streams. We performed experiments in Massive Online Analysis (MOA) environment. It is one of the most popular applications to work with data streams. It can both generate artificial data streams based on previous scientific works and also generate streaming data directly from stationary data sets. The primary requirement for the classifier model was the fact that it must have a minimal number of parameters and should be easy to understand. As a result, we chose the nearest neighbour classifier with non-standard distance measures. This work will present recent results.

Differences between STEM and Humanities & Social Sciences' Students: Evidence from Complex Tests of Visual-Spatial and Verbal Skills

L. Zariņa[1], J. Šķilters[1], G. Theara[2], K. Pētersone[1], N. Bērziņa[1]

[1] Laboratory for Perceptual and Cognitive Systems
Faculty of Computing, University of Latvia
[2] Georg-August-Universität Göttingen, Germany
liga.zarina@lu.lv

Links between the fields of study and spatial skills have been confirmed in several studies (e.g., Uttal et al, 2013, Wai et al, 2009). In our study, we focus on more specific kinds of spatial and non-spatial skills that we relate with the effects of demographic factors to determine the differences between STEM and Humanities/Social sciences' students. We tested Humanities/Social sciences (n=66) and STEM (n=49) students in a repeated measures quasi-experimental paper-pencil study containing six different tasks linked to different spatial, visual-spatial, and verbal skills: (1) Spatial orientation und rotation tests: (1.1.) mental rotation, (1.2.) spatial orientation; (2) Object imagery tests: (2.1.) visual imagery, (2.2.) snowy pictures (object recognition); (3) Verbal tasks: (3.1.) thing-categories, (3.2.) making sentences, and a comprehensive demographic part. In our study, STEM students indicate relatively better spatial skills in all tasks except mental imagery test (2.1.). Demographic differences indicate that male participants are better in spatial orientation (1.2.) and rotation tests (1.1.) while female participants show better results in making sentences (3.2.), image recognition (2.2.) and mental imagining tasks (2.1.). According to our results, a complex framework distinguishing between visual, visual-spatial, and verbal skills seems to be valid in determining differences between students of different fields; further, demographic variables seem to be sensitive set of indicators when exploring the differences between the fields of study.

Trends of Secondary Analysis of International Student Achievement Studies

R. Želvys[1], R. Dukynaitė[1], A. Jakaitienė[2]

[1] Vytautas Magnus University
[2] Vilnius University
rimantas.zelvys@fsf.vu.lt

International student achievement studies – PISA, TIMSS and PIRLS – are primarily focused on providing data on student achievement in comparative perspective. However, the surveys also contain information about the socioeconomic status of students, school systems, management and evaluation, etc. The collected data provides good opportunities for secondary analysis. The amount of publications of that kind is increasing each year. The aim of our presentation is to review main trends of secondary analysis in related publications.

On Some Old Problems of Bayesian Global Optimization

A. Žilinskas

Institute of Data Science and Digital Technologies
Vilnius University
antanas.zilinskas@mii.vu.lt

The interest of researchers to Bayesian global optimization is recently increased. One of resons of the increased interest is the new applications in machine learning. However, the broader applications are prevented by some problems which are still not solved although known for many years. In the talk we are going to discuss the following problems as well as the possible ways to cope with them. A crucial problem of the considered approach is the selection of a suitable statistical model with respect to the criteria of adequacy, estimability of parameters, and computational complexity. The second problem is rationality of search strategies. The third considered problem is that of the convergence.

The research was supported by Research Council of Lithuania under Grant No. P-MIP-17-61.

For notes

For notes

10th International Workshop

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

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

Vilnius University Press
1 Universiteto St., LT-01513 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt

Printed by "BMK leidykla"
16 J. Jasinskio St., LT-01112 Vilnius

Print run 150 copies