LITHUANIAN COMPUTER SOCIETY

VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES

LITHUANIAN ACADEMY OF SCIENCES

11th International Workshop on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

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

November 28–30, 2019

LITHUANIAN COMPUTER SOCIETY

VILNIUS UNIVERSITY
INSTITUTE OF DATA SCIENCE AND DIGITAL TECHNOLOGIES

LITHUANIAN ACADEMY OF SCIENCES

11th International Workshop on

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

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

November 28–30, 2019

VILNIUS UNIVERSITY PRESS

Vilnius, 2019

Co-Chairmen:

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

Programme Committee:

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

Organizing Committee:

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

Contacts:

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

https://doi.org/10.15388/Proceedings.2019.8

ISBN 978-609-07-0325-0 (digital PDF)

Preface

DAMSS-2019 is the 11th 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. History of the workshop starts from 2009 with 16 presentations. The idea of such workshop came up at the Institute of Mathematics and Informatics that now is the Institute of Data Science and Digital Technologies of Vilnius University. The Lithuanian Academy of Sciences and the Lithuanian Computer Society supported this idea. This idea got approval both in the Lithuanian research community and abroad. The number of this year presentations is 77. The number of registered participants is 127 from 9 countries. 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 9 companies and institutions supported the workshop this year. This means that the topics of the workshop are actual for business, too. Topics of the workshop cover Artificial Intelligence, Big data, Bioinformatics, Blockchain technologies, Business Rules Software Engineering, Data Science, Deep Learning, Digital Technologies, High-Performance Computing, Machine Learning, Medical Informatics, Modelling Educational Data, Ontological Engineering, Optimization in Data Science, Signal Processing, Visualization Methods for Multidimensional Data. A special session and discussions are organized about topical business problems that may be solved together with the research community. This book gives an overview of all presentations of DAMSS-2019.

Supported by:

General sponsors

Research Council of Lithuania
www.lmt.lt

Algoritmų sistemos
www.algoritmusistemos.lt

Main sponsors

Novian
novian.no

Western Union Processing Lithuania
www.westernunion.com

VTeX
vtex.lt

Sponsors

Baltic Amadeus
ba.lt

NetCode
netcode.lt

NRD CS
www.nrdcs.lt

Visoriai Information Technology Park (VITP)
vitp.lt

Sepsis Prediction Model Based on Vital Signs Related Features

Vytautas Abromavičius, Artūras Serackis

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

The presented investigation faces the problem of sepsis early detection. Our proposed algorithm uses three separate models for sepsis prediction. The current model is selected according to the time the patients have already spent in the intensive care unit. The first model uses 64 features and is applied if the patient stays in ICU for the first 9 hours. Second and third models use more advanced 111 features. The second prediction model is activated if the patient stays for more than 9 hours. The third prediction model is activated for more extended stays if the patient stays for more than 60 hours. For evaluation of our solution, we used data from PhysioNet/Computing in Cardiology Challenge 2019 test Set A and Set B. Inspection of the data, and consultation with personnel of the local ICU led us to believe that the time patients spent in the ICU or were hospitalized is an essential indicator for the risk of developing sepsis. During the longer stays in hospital number of intravenous measurements and other procedures increases, increasing the risk of blood infection. Therefore, feature extraction in our algorithm was based on these metrics. The performance of the proposed approach was tested using models trained with Gentle Boosting and alternative models, based on LSTM.

Do We Really Know How to Measure Software Quality?

Vineta Arnicāne, Juris Borzovs, Anete Kristīne Nesaule

University of Latvia
juris.borzovs@lu.lv

ISO/IEC 2502n series of standards that are devoted to system/software quality measurement currently consists of the following International Standards:

- ISO/IEC 25020 – Measurement reference model and guide: provides a reference model and guide for measuring the quality characteristics defined in ISO/IEC 2501n quality model division;
- ISO/IEC 25021 – Quality measure elements: provides a format for specifying quality measure elements and some examples of quality measure elements (QMEs) that can be used to construct software quality measures;
- ISO/IEC 25022 – Measurement of quality in use: provides measures including associated measurement functions for the quality characteristics in the quality in use model;
- ISO/IEC 25023 – Measurement of system and software product quality: provides measures including associated measurement functions for the quality characteristics in the product quality model.

Quality measure elements (e.g., number of faults) provide a base for measures of product and quality in use (e.g., functional correctness).

The presentation examines practical completeness of the ISO/IEC 2502n series software product part.

Multisensor Data Fusion for Data Analysis

Jaroslava Arsenjeva, Viktor Medvedev, Gintautas Dzemyda

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

Data fusion is the process of combining data from multiple sensors into one framework to provide better data analysis and improve decision making. It is a rapidly evolving trend among other ones such as IoT, Industry 4.0 and Big Data. And like every method, data fusion has its own difficulties. From having to deal with heterogeneous data and noise, different sampling rates and improper weight assignment for the raw data to receiving inferences that are contradictory – all of these are issues that can be assessed with the help of Kalman filter or Bayesian/ Demster-Shafer methods or some other well-known algorithms.

However, there are other complexities that have been known for more than 20 years and a lot of attempts were made to solve them however the ideal solution is yet to be found. The issues arising in data fusion implementation include the fact, that there is no ideal algorithm for any situation, a faulty sensor cannot be "replaced" by a complex framework, there is not enough sufficient training data due to changing environmental conditions and the value of output is hardly quantifiable. These problems possess a great interest in multiple fields and some solutions have been proposed however the determination of the best way to tackle them is yet to be determined.

The field which possesses the most promise for data fusion mining and analysis is the medical field. In the recent studies, a lot of attention is paid to human condition monitoring such as blood pressure, body temperature, heart rate and others. After processing the information from wireless sensors an accurate prediction about the overall state of the patient can be made.

Localization Algorithm for Identification of Mobile Objects in an Ad-Hoc Internet of Things Network

Kazimieras Bagdonas, Algimantas Venčkauskas

Kaunas University of Technology
kazimieras.bagdonas@ktu.lt

In this paper we present a localization algorithm for multimodal identification of mobile objects in an Internet of Things (IoT) network. Proposed algorithm is designed to aid data integrity verification and cyber security measures in a dynamic ad-hoc network, where localization with Global Positioning Satellite Systems (GNSS) is not available. IoT nodes are able to perform ranging measurements through communication channel and are able to estimate relative velocities between two communicating nodes. Obtained measurements are shared between neighboring nodes and used in resolution of the Local Reference Frames (LRF). Algorithm is able to propagate the change in the positions of IoT nodes in order to perform node identification and data verification tasks. Simulation of algorithm is analyzed and discussed. Obtained result verify the validity and scalability of proposed method. Novelty of proposed method consist of integration of velocity estimation to aid the resolution and propagation in time of LRF. Proposed algorithm accelerates the localization of an ad-hoc IoT network in decentralized fashion.

Computational Modeling, Multi-Objective Optimization and Decision Visualization of Microbioreactor System

Romas Baronas[1], Juozas Kulys[2], Linas Litvinas[1], Linas Petkevičius[1], Karolis Petrauskas[1], Antanas Žilinskas[3]

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

Bioreactors, based on microparticles containing immobilized enzyme, permit a specific substrates conversion to a certain product, a use of small volumes of samples and reagents, reduced costs, short processing time and system compactness.

This paper presents a Bayesian approach rooted algorithm oriented to the properties of multi-objective optimization problems. The performance of the developed algorithm is compared with several other multi-objective optimization algorithms. The approach is applied to the multi-objective optimization of a batch stirred tank reactor based on spherical catalyst microreactors.

The microbioreactors are computationally modeled by a two-compartment model based on reaction-diffusion equations containing a nonlinear term related to the Michaelis-Menten kinetics of the enzymatic reaction with addition of the mass transfer of the substrate outside the catalyst region.

A two-stage visualization procedure based on the multi-dimensional scaling is proposed and applied for the visualization of trade-off solutions and for the selection of favorable configurations of the bioreactor.

This research was funded by two grants (No. S-MIP-17-61 and No. S-MIP-17-98) from the Research Council of Lithuania.

Functional Data Analysis of fNIRS Data

Karolis Bartkus[1], Jurgita Markevičiūtė[1], Sigita Činčiūtė[2]

[1] Institute of Applied Mathematics
Vilnius University
[2] Institute of Biosciences, Life Sciences Center
Vilnius University
karolis359@gmail.com

Neuroscientists are interested by underlying functions in human brains and nerve systems. There are mainly two difficulties to organize groups of patients for such experiments. Firstly, the price of gathering the data. Secondly, the factors which influence the results has yet to be discovered. The fNIRS had been used as one of cheaper data gathering methods. In order to understand gender effect to fNIRS data functional ANOVA and MANOVA were used. By simulations the sample size and number of features are obtained to confident use of fMANOVA. The fMANOVA test is not stable but it could be a good indicator whether the samples have similar mean functions. The left frontal lobe has main differences between genders and that is in agreement with the knowledge of gender influence to experiments.

On Visualization of Properties and Structures Related to Zeta-Functions

Igoris Belovas[1,2], Martynas Sabaliauskas[2]

[1] Vilnius Gediminas Technical University
[2] Institute of Data Science and Digital Technologies
Vilnius University
igoris.belovas@vgtu.lt

Numerical calculation of zeta-functions, numerical investigation of their properties, as well as visualization, requires a significant amount of computations. In this research, we introduce a modification of Hasse's series representation for the Riemann zeta-function. We prove a limit theorem for the coefficients of the modified series and establish the rate of convergence to the limiting distribution. We use the result to construct an efficient algorithm for the calculation of the Riemann zeta-function. We discuss numerical aspects of the proposed technique and present our results on modified Hasse's series-based 3D visualizations, illustrating the underlying structure of surfaces and curves, associated with zeta-functions.

Scalability of Proof-of-Work Blockchains Using Off-Chain Solutions

Rytis Bieliauskas, Ernestas Filatovas, Remigijus Paulavičius

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

Current proof-of-work based blockchains have significant scalability limitations. The biggest and most widely used proof-of-work blockchain (Bitcoin) allows the network to process 5-10 transactions per second. Some other proof-of-work blockchains theoretically have a higher transaction per second throughput. Still, there are no other proof-of-work based blockchains where a higher number of transactions would be happening constantly and reliably. An off-chain solution, called the Lightning Network, enables processing of thousands of times more transactions per second. Additionally, the Lightning Network decreases transaction confirmation time from 10-60 minutes to less than a second, significantly reduces transaction fees and increases the privacy of the users.

Mobile Platform for Fatigue and Workability Evaluation: Functional State Data Visualization

Liepa Bikulčienė, Tomas Blažauskas, Aušra Žvironienė, Kristina Poderienė

Kaunas University of Technology
liepa.bikulciene@ktu.lt

Human functional state is one of the most important factors that determines his working capacity. Working conditions usually are not very convenient (especially in manufacturing, construction, services, etc.). Therefore, it is important to determine the level of work capacity and fatigue of the employee in order to avoid accidents or injuries during work. It can be measured quickly enough with our proposed methods and algorithms and can advise a person how to act in the future to prevent health problems or what signs of illness to look out for. Also, a big part of population is suffering from fatigue, which reduces the work capacity and quality of life. The information about Eureka project Fatigue - "Non-intrusive human fatigue assessment" will be introduced in this presentation. The aim of this project is to develop a platform with a 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. Lithuanian partners will develop a fatigue monitoring system, based on the use of unobtrusive sensors, interactive questionnaires and best signal quality during daily activities. An 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 daily-life activities. Results of interactive questionnaires and physical tests must be understandable to the participants themselves and clearly identify problematic health areas. For this reason, a visualization similar to wind roses was proposed. This will be included into software modules for data processing and feedback, i.e. providing recommendations for different use cases.

This research was funded by a grant (No. 1.2.2-MITA-K-702) from the Agency for Science, Innovation and Technology (MITA) regarding Eureka project "Non-intrusive human fatigue assessment" (Eureka 11169 Fatigue).

Narrative Detection for Lithuanian Language

Monika Briedienė[1,2], Tomas Krilavičius[1,2]

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

Automatic narrative detection is an important problem in text analytics, covering a number of different methods and applications. Due to growing amount of fake news, propaganda and, in general, growth of different media and media channels, interactive narratives analysis become essential.

We analyze a development of research on the topic over 2013-2018 years with different applications, methods, evaluation metrics, experimental corpora and maturity for different languages and tasks. When analyzing articles by selected keywords it was noticed that the task of automatic narrative (or its parts) detection is very interdisciplinary and most often solved not only in informatics but also in philology, medicine, social and political sciences. The studies discussed tend to analyze specific cases with specific language corpora (e.g. *Analysis and Automatic Classification of Some Discourse Particles on a Large Set of French Spoken Corpora; Topics of Ethnic Discussions in Russian Social Media; Memory, Party Politics, and Post-Transition Space: The Case of Poland; Automatic deception detection in Italian court cases*, etc.). Most of the investigation was done with English language and no multi-language cases were noticed. Data sets, that are usually used in experiments, are of two types: formal (normative) and informal (non-normative) texts; most commonly used media channel was Twitter (about 60 % of all media cannels).

Due to the problem complexity, we look at different sub problems, namely – crawling, classification, segmentation, sentiment analysis, event detection and others, their quality and maturity. Our goal is the development of methodology for narrative development analysis (with Lithuanian corpora), which allows following dynamics of narrative, related sentiments and events, differences of narrative from different information channels, changes over time.

In order to solve the problem of automatic narrative detection in the Lithuanian media, it is first necessary to collect proper (annotated) corpus. Then text preprocessing is needed, with appropriate data machine learning approaches could be used. The purpose of these operations is to construct a mapping between the elements of narrative structure and the discourse relation classes. Once a common narrative detection model has been constructed, it will be able to solve a variety of relevant cases.

Fractal Dimensionality of Network Traffic as a Feature for Intrusion Detection

Viktoras Bulavas

Institute of Data Science and Digital Technologies
Vilnius University
viktoras.bulavas@itpc.vu.lt

Cyber threats are an evolving aspect of our daily lives, intrusion detection being one of the remedies to address information security breach. Intrusion detection relies on observation of network traffic features and their dynamics in time, which allows intrusion detection systems to prevent certain types of attacks upon detection. While rule based systems are following decision trees of prescribed conditions, anomaly recognition systems await for deviation from usual behavior of network users. While multiple event counters help rule-based recognition, various aggregates are calculated in order to detect anomalies. Based on the analogy of successful use of fractal features to recognize patterns in other fields of application including signal analysis, an experiment with network traffic dataset CSE-CIC-IDS2018 was setup to study fractal dimension features of network traffic as a possible indication of a cyber-attack. Network traffic aggregates were represented as two-dimensional images, further presented as an animation, allowing real time observation of the development of an attack. To support cyber-attack detection, Box-Counting calculation according to T. Higuchi algorithm was performed to extract fractal dimension of a given timeframe traffic block. Maximum values were observed at the time of an attack and minimum following the successful attack. These results are in line with dataset events, confirming a possibility to use this feature for supporting real time detection of cyber-attack.

Approximating the Minimum of a Smooth Gaussian Process

James M. Calvin

Department of Computer Science
New Jersey Institute of Technology, USA
james.m.calvin@njit.edu

Many of the difficult questions concerning global optimisation are interesting and difficult even in the one-dimensional case, where they are more easily described. The purpose of this talk is to explore some of the natural questions that arise in the study of the average-case complexity of global optimisation of smooth Gaussian processes. The reason for studying the average-case complexity of optimisation is that if one optimises over a convex and symmetric function class, nonadaptive algorithms are essentially as powerful as adaptive methods. For example, if the goal is to approximate the minimum of a function defined on the unit interval that is only known to be twice continuously differentiable, then evaluating the function at equi-spaced points is near optimal in terms of the worst-case error. In practice, adaptive methods are favoured, and while they cannot be justified by their worst-case performance, adaptive algorithms have been shown to be much more efficient than nonadaptive methods on the average for some probability models. In this talk, we will consider the following type of question: Given an error tolerance $\epsilon > 0$, on average how many evaluations of the function are required to obtain an expected error of at most ϵ? We will examine how the answer depends on the information available to the algorithm as well as the probability model.

Expanding Convolutional Networks with SIFT Features to Classify Images Better

Bernardas Čiapas, Povilas Treigys

Institute of Data Science and Digital Technologies
Vilnius University
bernardas.ciapas@mif.vu.lt

Convolutional neural networks (CNNs) yield state of the art image classification results, yet these results are insufficient for many computer-vision based applications. Creating well performing CNNs comes with a price: they require many labelled images for training, which is expensive to obtain in many industries; image features extracted by convolutions are location dependent.

Image features obtained using various other methods than convolutions – for example, Scale Invariant Feature Transform (SIFT) – are different by nature, therefore, complementary. Combined features obtained from convolutions and SIFT carry more information than from convolutions alone.

In this research we will combine CNNs with image features obtained using other techniques (e.g. SIFT). We hypothesize that combining CNN and SIFT features will yield better accuracy than using CNN alone. Since SIFT appetite for data is lower than CNN, we make another hypothesis: when amount of training data decreases, gap in accuracy between CNN and SIFT+CNN increases.

New Applications of Sound and Vision Engineering to Information and Communication Technology

Andrzej Czyżewski

Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology, Poland
andcz@sound.eti.pg.gda.pl

Sound & Vision Engineering is being explored and taught at Gdansk University of Technology (Gdansk, Poland) for nearly 5 decades. The scope of scientific interests of the department covers many topics, including: multimedia technology, digital signal and image processing, particularly based on methods pertaining to the field of artificial intelligence and telecommunications, speech acoustics, the psychophysiology of perception, advanced image processing with applications in biomedical engineering, biometrics, public safety, and also cultural heritage restoration. In recent years, new applications for automated analysis and processing of signals, image and video data are being developed which will be discussed in the keynote paper illustrate the scope of the currently carried-out research in the department, mainly in the domain of video and sound processing and their applications. Some recent project results will be shown in order to illustrate the progress made in this discipline. For example, the project ALOFON (Audiovisual speech transcription method) is to conduct research aimed at developing a methodology of automatic speech phonetic transcription (in English), based on the use of a combination of information derived from the analysis of audio, video and facial motion capture signals. The project HCIBRAIN (Human-computer communication methods for diagnosis and stimulation of patients with severe brain injuries) develops in accordance with the assumption regarding the implementation of basic research and experiments in the field of diagnosis and therapy of non-communicating patients. An integrated multimodal system was developed together with a diagnostic and therapeutic procedure to the diagnosis and rehabilitation of severely impaired patients, in particular, those remaining in a coma. The project IDENT (Multimodal biometric system for bank client identity verification) the team has built a scientific synergy with the biggest Polish Bank (PKO BP), both in terms of technical cooperation, as well as in the domain of assessing the feasibility of implemented biometric solutions which provide the subject of joint research and development work finished within 2018. The objective of the project INZNAK (Intelligent Road Signs with V2X Interface for Adaptive Traffic Controlling) is to develop a conceptual design and research tests of a new kind of intelligent road signs which will enable the prevention of the most common collisions on highways, resulting from the rapid stacking of vehicles resulting from accidental heavy braking. The most recent project, namely project INUSER (Integrated Systems for Managing Wind Farms) assumes development of a set of solutions for the monitoring and the diagnosing the condition of selected parts of power wind turbines. The spatial distribution of vibroacoustic energy and its propagation based on the measurement of parameters of sound intensity within given gridpoints, employing a special probe developed by the department's research & engineering team is the main challenge of this newest project.

Application of Deep Learning Methods in Host-Based Intrusion Detection Systems

Dainius Čeponis, Nikolaj Goranin

Vilnius Gediminas Technical University
dainius.ceponis@vgtu.lt

Protection of information plays an important role in the daily schedule of a modern company. Various types of businesses are dealing with a huge amount of sensitive data: it can be not only data belonging to the private company but also personal data of employees or customers' information. Intrusion detection systems (IDS) are used to prevent events when malicious third parties seek to gain access to critical information. Early implementations of IDS systems had simple decision-making engines and used a trivial amount of data, including known attack patterns and were useless against zero-day attacks. More extensive operations have to be executed by the IDS today. Various machine learning (ML) models are proposed to be used for these tasks. They demonstrate high detection rate and small false positives when deciding is any action is intrusion or not. Convolutional Neural Networks, Recurrent Neural Networks and LSTM (Long Short-Term Memory) Networks are among the most advanced ML methods. They can automatically extract important features from the data and perform an accurate attack classification. Classification effectiveness of all listed methods has been tested on Windows OS generated System-Calls data, collected in a newly created AWSCTD dataset. The achieved results demonstrate deep learning methods can be successfully used for intrusion detection on the Host level with up to 95% accuracy.

Study of Blockchain-Based Smart Contract Application for Peer-to-Peer Electricity Exchange in Distributed Trading Network

Paulius Danielius, Saulius Masteika

Kaunas Faculty
Vilnius University
saulius.masteika@knf.vu.lt

In the perspective of requirements for future energy systems – decarbonisation, decentralization and digitalization, and increasing deployment of renewable energy sources, the transformation of energy distribution and trading mechanisms is inevitable. As the number of households are acquiring independent power generation capacities and becoming prosumers willing to sell excess energy, while energy consumers are looking for fair prices, the need for more flexible and cost-effective way to exchange energy is rising. Blockchain technology is designed to facilitate distributed transactions by removing the need for central management and utility-intermediaries, and as such is identified as potentially suitable platform for distributed energy market. While interested parties are investigating feasibility of using blockchain technology in energy sector, and number of pilots and research projects grows, details on realization are usually lacking. In our work, we are investigating smart contract application for peer-to-peer energy exchange in distributed blockchain-backed platform.

A Study of Supervised Combined Neural-Network-Based Ultrasonic Method for Reconstruction of Spatial Distribution of Material Properties

Paulius Dapkus

Kaunas University of Technology
paulius.dapkus@gmail.com

This paper examines the performance of the commonly used neural-network-based classifiers for investigating a structural noise in metals as grain size estimation. The biggest problem is to identify the object structure grain size based on metal features or the object structure itself. When the structure data is obtained, a proposed feature extraction method is used to extract the feature of the object. Afterwards, the extracted features are used as the inputs for the classifiers. This research study is focused to use basic ultrasonic sensors to obtain objects structural grain size which are used in neural network. The performance for used neural-network-based classifier is evaluated based on recognition accuracy for an individual object. Also, traditional neural networks, namely convolutions and fully connected dense networks, are shown as a result of a grain size estimation model. To evaluate robustness property of neural networks, the original samples data is mixed for three types of grain sizes. Experimental results show that combined convolutions and fully connected dense neural networks with classifiers outperform the others single neural networks with original samples with high SN data. The dense neural network as itself demonstrates the best robustness property when the object samples do not differ from trained datasets.

Overview of Identification of Phishing Email Messages

Laurynas Dovydaitis, Kęstutis Driaunys

Kaunas Faculty
Vilnius University
laurynas.dovydaitis@knf.vu.lt

Phishing social engineering attack is targeted towards individuals and organizations. Typically, the attack is conducted in two steps, where an email is sent to a victim and then an attacker waits for the victim to follow an URL link embedded in that email message. This overview provides the identification of phishing email messages. After analyzing the differences between the two methods, it is proposed to combine two classification feature sets as a new approach to identifying a phishing message.

Online Estimation of Parameters for Discrete-Time Independent Normal Random Variables Summation Process Observed with Noise

Vytautas Dulskis, Leonidas Sakalauskas

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

The linear dynamical system (LDS) model is used ubiquitously in practical applications. However, the problem of system identification for the LDS model is currently better solved offline than online, which suggests that further improvements need to be made in terms of solving such a problem in the online case. In this work, we propose an algorithm for the online estimation of parameters for discrete-time independent normal random variables summation process observed with noise, which is a special case of the LDS model. The algorithm uses maximum likelihood estimation in combination with a certain parametrization of unknown model parameters and Taylor series approximation, and may provide insight into the online solution of the system identification problem for the general LDS model.

New Ideas for Multidimensional Scaling

Gintautas Dzemyda, Martynas Sabaliauskas

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

Multidimensional scaling (MDS) is the most often used method for dimensionality reduction of data. A novel geometric interpretation of the multidimensional scaling process (Geometric MDS) is discussed. MDS is based on the minimization of so-called stress function dependent on the coordinates of data points in lower dimensionality. According to the new interpretation, attention is paid to the local stress function dependent on the separate lower-dimensional point. The proposed geometric step guarantees the decrease of the local and global stress by recalculating coordinates of this point. A consequent selection of low-dimensional points and their movements to a new position using a new approach (Geometric MDS) guarantees to reach at least the local minimum of the global stress function. This method was compared with SMACOF 2.0 version. The experimental results showed that both of these methods have a similar efficiency when the obtained stress values are compared. The used realization of Geometric MDS works slower than SMACOF 2.0. However, Geometric MDS is much simpler to realize and comprehend.

Acceleration of ISOMAP for Hyperspectral Image Classification on Multicore Processors and GPUs

Ernestas Filatovas[1], Francisco Orts[2], Gloria Ortega[3], Olga Kurasova[1], Ester Martín Garzón[2]

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

The isometric mapping (Isomap) is a nonlinear dimensionality reduction method that is often used for analyzing hyperspectral images. To achieve such objective, Isomap involves the construction of a neighbourhood graph and the computation of the shortest paths between the nodes. Moreover, it uses the state-of-the-art MultiDimensional Scaling method (MDS) for dimensionality reduction. In this work, two efficient parallel versions of ISOMAP using OpenMP and CUDA have been developed and evaluated on multicore and GPU. On the one hand, we propose to use Isomap with SMACOF, the most accurate MDS method. On the other hand, we propose to use an insertion sort instead of the k-nearest neighbours' algorithm (KNN) for the construction of the neighbourhood graph. Since the number of managed neighbours by ISOMAP is low, the insertion sort is probed to be more efficient than KNN.

State Space Representation of Lee-Carter Stochastic Mortality Model: Application to Modelling of Insurers' Solvency Capital

Rokas Gylys, Jonas Šiaulys

Vilnius University
rokas.gylys@mif.vu.lt

In the presentation we show how state space methods can be used as an alternative fitting procedure for Lee-Carter stochastic mortality model. Such modelling approach substantially increases model's flexibility and, in contrast to the two-stage estimation strategy of the classic Lee-Carter, ensures a coherent statistical estimation of the parameters. We also show how Kalman filter in combination with Gibbs sampler can substantially simplify Bayesian data analysis for such models. The application of the methods is illustrated with the results of calculations on the mortality data of Lithuania and Sweden using both the classic Lee-Carter model specification and its extensions. Finally, we demonstrate how the results of the calculations are used to perform the assessment of the solvency capital of insurance companies.

Habits Attribution & Digital Evidence Object Tool with Fuzzy Logic for Cybercrime Investigation

Šarūnas Grigaliūnas, Jevgenijus Toldinas, Borisas Lozinskis

Kaunas University of Technology
sarunas.grigaliunas@ktu.lt

In cybercrime investigation, methods and tools play an important role in extracting the evidence from various digital places. Theoretical methodologies define models to investigate cybercrimes, and practical tools provide abilities for investigators to extract the evidence and prove the crime. A habits attribution method is used to identify habits, attribute them, and then create a profile of the attributed habits. The created profile, as set of habits and attributes, is used in digital evidence investigation to reduce the number of evidence search sequences from a set of the digital artifacts, where each artifact is based on the user habits attribution. A digital evidence object is defined as DEO = (Why, When, Where, What, Who). The DEO model is used for forensics data reduction in digital forensic investigation, and in conjunction with the habits attribution model is realized in the proposed tool. Fuzzy logic at approximate reasoning is effective in situations where information is ambiguous, insufficient or inaccurate. The proposed tool in which fuzzy logic is applied to the habits attribution and digital evidence object models helps investigators in making decision to extract the evidence and prove the crime. We present a real-world case study of the proposed tool to demonstrate its applicability for assisting investigators in the digital evidence investigation process. Our results show that fuzzy logic application in conjunction with the habits attribution and digital evidence object models can significantly reduce the number of false positive evidence objects presented to an investigator, reduce investigator workload and improve decision making performance.

Aligning Agile Application Software Development with the Archimate Framework

Saulius Gudas, Karolis Noreika

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

The effectiveness of internal business processes is a key in the modern-day economy for business enterprises of all sizes. This includes the enterprise application software (EAS) development management and its alignment with organizational goals both short and long term. This is a difficult question of how to balance EAS development with organizational goals.

This paper suggests the principles for aligning Agile software development approaches with enterprise architecture framework ArchiMate 3.0. Enterprise business managers are also starting to make decisions based on Agile methodology, although it is often perceived as a part of startup culture. Often, new EAS developments are ahead of new business solutions where application software development teams need rapid deployment related to business process changes. In addition, these new EAS developments are aimed at reusing in multiple projects in the same business area. By aligning business needs with software changes, the Agile development methodology could be applied to business decision making, linking Agile life cycle phases to the decision making steps.

Agile software development methods provide the speed and the possibility to adapt to changes and should be used with enterprise architecture frameworks like ArchiMate, which provides the tools to ensure business strategy and processes alignment to EAS architecture, to make the most benefit of both of the approaches. This approach allows us not only to visualize the Agile software development process steps and thus obtain specifications for the EAS project, but also to keep EAS projects consistent with business strategy.

Experimental Investigation of Energy Consumption for Cryptocurrency Mining

Aleksandr Igumenov, Ernestas Filatovas,
Remigijus Paulavičius

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

Bitcoin was introduced in 2008 by an anonymous person or group of individuals called ‘Satoshi Nakamoto’ to work as the public ledger of transactions, forming a chain of blocks, hence “blockchain”. The first and the most popular applied usage of blockchain technology is cryptocurrencies like Bitcoin, Litecoin, Monero, etc. For the generation of the new cryptocurrencies (“mining”), different computational resources are used: computer central processing units (CPU), graphics processing units (GPU), specific integrated circuits (ASIC). Hardware is used to solve a resource-intensive cryptographic hashing problem — “Proof of Work (PoW)”. This mining process contributes to blockchain network security, however, it can be very greedy in electricity. In this study, we present an investigation of mining efficiency for various cryptocurrencies with different devices.

Gearing-up Cooperative Multiobjective Optimization

Ignacy Kaliszewski, Janusz Miroforidis

Systems Research Institute of the Polish Academy of Sciences
ignacy.kaliszewski@ibspan.waw.pl

Here we deal with large-scale multiobjective optimization problems of sizes that premium commercial solvers get stuck with the memory or time limit. In our recent work we have shown that solving multiobjective optimization problems to suboptimality (because of the limits) in a cooperative manner improves lower and upper bounds on components of elements of the Pareto front. By cooperative solving we mean the exchange of information gathered when solving for individual Pareto optimal solutions, for the benefit of the cooperating pool.

In this presentation we shall show how cooperative multiobjective optimization can be geared-up by exploiting specific properties of combinatorial multiobjective optimization problems and their surrogate constraint relaxations. We illustrate our developments by numerical experiments on large-scale multidimensional bicriteria knapsack problems.

Predictor-Based Self-Tuning Control of Emotion Signals as Response to a Dynamic Virtual 3D Face

Vytautas Kaminskas, Edgaras Ščiglinskas

Vytautas Magnus University
vytautas.kaminskas@vdu.lt

This paper introduces the application of self-tuning control with constraints of human response to a dynamic virtual 3D face. We are using 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 or frustration signal – output signal. The technique of on-line identification which ensures stability and possible higher gain for building a predictive input-output models is applied. Predictor-based self-tuning control schemes with a minimum variance and generalized minimum variance controllers with constrained control signal magnitude and change speed are developed. Analysis of prediction-based self-tuning control results demonstrate sufficiently good control quality of excitement and frustration signals. Stabilized excitement or frustration signal level is on average higher compared to the average of observed response as reaction to the testing input. Experiment results of the control demonstrated possibility to decrease variations of the virtual 3D face using a limited control signal change speed. Variation of the virtual 3D face at the same limited speed values is higher in the self-tuning control scheme with a minimum variance controller compared to variations in the schemes with generalized minimum variance controllers.

Towards Creation of the Deep Learning-Based Natural Language Understanding Module for the Lithuanian Chatbots

Jurgita Kapočiūtė-Dzikienė[1,2], Kaspars Balodis[3],
Raivis Skadiņš[3]

[1] JSC "Tilde informacinės technologijos", Vilnius, Lithuania
[2] Vytautas Magnus University
[3] JSC "Tilde", Riga, Latvia
jurgita.k.dz@gmail.com

Natural Language Understanding (NLU) is responsible for the comprehension of meaning and structures in the human language. In recent years, traditional machine learning approaches in the NLU module are replaced with the Deep Learning (DL) techniques. Despite that, there is no one right solution that fits the best for all languages: methods require exploration and adaptation to each language specifics. The Lithuanian language is much more challenging compared to the highly-explored and resource-rich English: Lithuanian is highly inflective, has rich vocabulary, word derivation system and relatively free word-order in the sentence. In this research we have solved the intent detection problem by using three benchmark conversational datasets: askUbuntu (with 5 intents), chatBot (2 intents), and webApps (8 intents) that were translated into the Lithuanian language. Different DL variants –in particular, Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) method and Bidirectional Long Short-Term Memory (BiLSTM) method– were experimentally explored. All DL techniques were applied on the top of two types of distributional word embeddings: the Lithuanian language-dependent fastText embeddings and language-independent BERT embeddings. Apart from the correctly selected DL method variant and vectorization, DL hyperparameters (such as a number of hidden layers, number of neurons, batch size, number of epochs, activation functions, optimization algorithm, dropout, recurrent dropout, etc.) also play an important role in boosting the overall intent detection accuracy. Hyper-parameters were automatically tuned using two optimisation algorithms: tpe.suggest (directional tuning) and tpe.random (random tuning). The CNN method demonstrated the best performance on the askUbuntu and webApps datasets, LSTM – on the chatBot dataset. The best results for all three datasets, i.e., askUbuntu, webApps and chatBot were achieved with the BERT embeddings reaching 0.881, 0.981 and 0.661 of the accuracy, respectively. Our offered techniques on the askUbuntu and webApps datasets outperformed the other popular Wit.ai and Microsoft LUIS chatbot systems.

Scheduling Complex Applications in Cloud Systems – Challenges and Research Directions

Helen Karatza

Department of Informatics
Aristotle University of Thessaloniki, Greece
karatza@csd.auth.gr

For several years now, cloud computing has become a popular computing model and has significantly impacted the IT sector. It offers almost unlimited computing resources to end-users for running complex computationally intensive applications. With the advent of the cloud computing paradigm, many new challenges and opportunities have appeared. However, there are important issues that must be addressed, in order to exploit cloud computing capabilities. This is due to the large scale of the cloud and the continuously increasing number of cloud users and complex applications deployed in it. Therefore, important research has been carried out in cloud computing which includes many areas such as resource allocation, scheduling, availability, cost, reliability, quality of service, energy conservation, virtualization. Efficient scheduling algorithms play a crucial role in cloud computing and must provide a good performance to leasing cost ratio. Very important in cloud computing is the effective scheduling of complex real-time applications, considering not only the processing times but also the cost of energy consumption. Therefore, energy-efficient scheduling strategies are required allowing for guarantees that the deadlines will be met. The goal of this talk is to present recent advances in cloud computing covering various concepts on complex applications scheduling and to provide research directions in the cloud computing area.

Fraudulent Transaction Path Detection in Ethereum Blockchain

Gabrielė Kasputytė[1,2], Karolis Lašas[1,2],
Rūta Užupytė[1,2], Tomas Krilavičius[1,2]

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

Cryptocurrencies are getting more and more popular as a medium of exchange while slowly pushing the standard currency out of the market. The development of the blockchain, currency's decentralization and the principle of publicity of transactions led to the success of cryptocurrencies. However, due to the novelty and minimal or absent regulation of the cryptocurrencies ecosystem, it attracts a high number of fraudulent activities. One of such activities is cloaking of incoming funds by using a number of intermediate wallets for a transaction. In this research we investigate an approach for such transactions detection in Ethereum ecosystem, namely, Ether transactions. Number of daily Ether transactions exceeds 500 000, hence detection of any fraudulent behavior is not trivial. We define all Ether transactions as a graph, where nodes are wallets and arcs correspond to transaction, while weights of nodes denote amount of Ether transferred from one wallet to another. Our proposed algorithm explores whole graphs, and checks each node and transaction, i.e. flow of Ether in the network. Clustering is applied to improve the accuracy of the algorithm. As a result, the algorithm builds a transaction tree visualizing Ether flow from wallet to wallet, namely transaction path. Moreover, it allows identifying relations of wallets to fraudulent wallets.

Discrete-Valued Time Series Analysis: Models, Methods and Software

Yuriy Kharin

Research Institute for Applied problems of Mathematics & Informatics
Belarusian State University
Kharin@bsu.by

Time series analysis is deep developed for “continuous” data when the observation space A is some Euclidean space or its subspace of nonzero Lebesque measure mes(A) > 0. In practice, however, (because of “digitalization” of our real world) the researchers need to use discrete-valued models of time series, when the observation space A is some discrete set with cardinality N = |A|, mes(A) = 0. Give some applied areas where discrete-valued time series models are extremely helpful: bioinformatics for analysis of genetic sequences (N = 4); information systems for information protection (N = 2); meteorology for weather prediction; social science for modeling of dynamics in social behavior; public health and personalized medicine; prediction of environmental processes; financial engineering; telecommunications; alarm systems.

We develop models, methods and software for computer analysis of discrete-valued time series. Two approaches to construction of parsimonious (small-parametric) models for observed discrete-valued time series are proposed based on high-order Markov chains.

A family of parsimonious models for discrete-valued time series are constructed. Consistent statistical estimators for parameters of the proposed models and some known models, statistical tests on the values of parameters, and also forecasting statistics are constructed. Probabilistic properties of the constructed statistical inferences are given. The developed methods are also applied for statistical analysis of spatio-temporal data.

Theoretical results are illustrated by results of computer experiments on real statistical data.

Receptive Field in Neural Network Keyword Spotting Models

Aliaksei Kolesau

Vilnius Gediminas Technical University
kolesov93@gmail.com

Many keyword spotting models use neural networks to detect acoustic events such as phonemes, word pieces or whole words. The model is inferenced on every frame (segmented piece of audio) which is typically every 10ms. In order to improve the quality of classification neural network uses audio features for both the frame under classification and several adjacent frames. This introduces a tradeoff. Too large receptive field might cause overfitting, increases the number of parameters and latency. Too small receptive field might not be able to provide enough information to correctly classify audio event. We investigate several policies of constructing receptive field for neural network in keyword spotting including the ways to make receptive field more sparse such as frame skipping and frame stacking.

Evaluation of Lombard Speech Models in the Context of Speech Enhancement

Gražina Korvel[1], Krzysztof Kąkol[2], Bożena Kostek[2]

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

The environment noise changes the manner of expression. The Lombard effect is one of the most known effects of noise on speech production. The results obtained in our previous study lead us to conclude that speech with the Lombard effect is more recognizable in noisy environments than normal speech. Our investigations have also shown that that speech synthesis model may retain Lombard effect characteristics. In this study, we investigate several models of Lombard speech in the context of speech enhancement. For this purpose, 25 statements (15 sentences and 10 words) uttered by four speakers were used. These statements were recorded in two conditions: without additional noise as well as with interference. These conditions resulted in two types of recordings: 100 statements of normal speech and 100 with the Lombard effect, i.e., non-Lombard speech. In the experimental part of the research, the Lombard speech models such as harmonic, source-filter, and these based on sinewave oscillator bank were investigated. The main goal was to check how these models are recognizable when the signal is reverberant and at what the noise threshold the model stops working. For this purpose, the models and Lombard speech were mixed with babble speech and street noise recordings with a different signal to noise ratio (SNR). The quality of these models was measured employing objective indicators. The experimental investigations show the superiority of source-filter models over other models utilized.

Signal Processing Methods Based on Multivariate Analysis for Wearable Devices Detecting Life-Threatening Health Conditions in Hemodialysis Patients

Algimantas Kriščiukaitis[1,2], Ana Rita Alves dos Santos Rodrigues[1], Andrius Petrėnas[1], Vaidotas Marozas[1]

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

Sudden cardiac death (SCD) accounts for a substantial amount of all deaths in end-stage renal disease (ESRD) patients due to arrhythmia and cardiac arrest. The exact mechanism by which SCD occurs in ESRD patients remains unclear since traditional risk factors for coronary artery disease are found only weakly associated with SCD in the dialysis population. The incidence of arrhythmias is notably higher during and within the hours before and following hemodialysis sessions. It suggests that anomalous serum electrolyte levels and the abrupt correction of such levels during hemodialysis significantly contribute to SCD. Patients with impaired kidney function are predisposed to develop metabolic acidosis and hyperkalemia (high potassium levels), which can induce a negative inotropic effect in the heart by altering its electrical activity and cardiac repolarization, leading to cardiac alternans or even evoking severe post-acidosis/hyperkalemia arrhythmias. Monitoring electrical activity of the heart could not only enable the detection of threatening conditions but elucidate the causes of such conditions in ESRD patients. The current rapid development of wearable devices offers the possibility to unobtrusively evaluate and monitor several physiological parameters, which can be reflective of the aforementioned critical pathogenic factors. In this work, we present methods for evaluation of alterations in cardiac repolarization, related to alterations in serum potassium concentration and metabolic acidosis, focusing on a detailed ECG T-wave analysis. Mul-

tivariate signal analysis methods (e.g. Principal Component Analysis) are used to evaluate variations in QRS- and T-wave shape and changes in spatial positions of cardiac depolarization and repolarization front vectors. To have a comprehensive representation of the heart's electrical activity, we start from analysis of multilead ECG recordings, aiming to make it suitable for wearable devices in the future by optimizing/minimizing the number of registered/analyzed signals.

Cognitive Mapping Principle for Advanced Survey Analysis

Dalia Kriksciuniene[1], Virgilijus Sakalauskas[1], Dale Luksaite[2]

[1] Kaunas Faculty, Vilnius university
[2] Kauno kolegija / University of Applied Sciences
dalia.kriksciuniene@knf.vu.lt

Evaluation of opinions and attitudes is one of the areas where the survey methods are prevailing. However, the problem of causal impacts within the set of questions of the survey and the goal pursued by the entire research cannot be solved by applying standard statistical approaches. The integration of principles of cognitive mapping and DEMATEL methodology is applied for uncovering hidden relationships among the concepts and evaluating their impact. The experimental research of the educational institution was performed for validating the proposed methodology.

Application of Deep Learning for Credit Scoring

Dovilė Kuizinienė[1,2], Tomas Krilavičius[1,2]

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

High competition among financial institutions and current automation trends requires automation of many banking processes, including credit scoring. One of the potential venues to improve credit scoring quality is the application of Artificial Intelligence methods. Usually, credit scoring is defined as a classification problem dividing potential debtors/loans into good (orderly paying their loans) and bad (defaulting) classes. We apply deep learning methods to credit scoring in this research, using three real World data sets from the UCI repository (German, Australian and Japan). The attributes from Australian and Japan data sets are unknown, it means that attribute names and values have been changed to meaningless symbols to protect the confidentiality of the data, hence we have to be careful while applying any transformation and feature engineering methods. We apply Convolution Neural Networks (ConvNet). Usually, it shows good accuracy results while working with high number of features, extracted and learned directly from the data. We present the following results: 1) analysis of random convent architectures for each data set, based on the logic of the wrapper method (to optimize the accuracy); 2) the best ConvNet architectures results compared to other machine learning methods; 3) the best ConvNet architectures results compared to results presented in other papers.

Application of Neural Networks for Cyber Attacks Detection

Rimantė Kunickaitė[1], Tomas Krilavičius[1,2]

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

Digital networks play an important role in processing, data storage, monitoring and control of critical objects of infrastructure such as energy, transportation, and finance. Due to the increasing number of cyber attacks and their complexity, critical computer networks become increasingly vulnerable. Disruption of these networks can have extremely harmful effects. Current cyber security systems are not able to protect from all attacks not providing near real time response. Host-based intrusion detection systems cannot protect enough these networks due to the sheer volume, distributed nature of data, and real-time response requirements. Furthermore, they only identify known attacks. During analysis of network traffic data, signature based and anomaly based intrusion detection systems enable to detect unseen attack and can verify data near real time. In this research, we investigate the applicability of Artificial Neural Networks for signature-based and anomaly-based cyber-attacks detection. The accuracy of selected models in the experimental research reaches up-to 97% for certain attacks. The results of research show the effectiveness of artificial neural networks in different cyber attacks identifying. Furthermore, we present future research plans, such as analysis of different cyber-attacks types and application of different methods, including ensemble approaches, as well as data augmentation.

A Method of Automatic Design of an Acoustic Diffuser Based on a Genetic Algorithm

Adam Kurowski, Bożena Kostek

Multimedia Systems Department and Audio Acoustics Laboratory
Faculty of Electronics, Telecommunications and Informatics
Gdansk University of Technology, Poland
bozenka@sound.eti.pg.gda.pl

A method of automatic design of an acoustic diffuser based on a genetic algorithm is proposed. This allows defining the geometry of an acoustic diffuser, which is tailored to the room of specific geometry and the purpose of music mixing or mastering music or performing other types of audio-related work. Parameters that are commonly employed to describe the quality of acoustic treatment, such as the size of a reflection-free zone, sound clarity and the ratio of direct to reverberant sound are employed as a fitness function for the genetic algorithm. They are computed with the use of the finite time difference method, which imposes relatively low requirements related to the computational power of a computer used for calculations. The experiment is conducted for selected geometries of rooms.

Optionally, a room designed also contains devices providing attenuation of unwanted sound reflections. The performance of Schroeder diffuser is compared with the performance of two reference designs obtained with the use of QRD (Quadratic Residue Diffusers) and PRD (Primitive Root Diffusers) pseudo-random sequences. The simulation is carried out for each acoustic diffuser designed. It is repeated several times with changes in the position of sound sources to estimate the quality of each design concerning the position of loudspeakers. The performance of diffuser geometries is compared for each of the four parameters with the use of the ANOVA statistical test and visualized in the form of boxplots.

Ranking-Based Discrete Optimization Algorithm for Asymmetric Competitive Facility Location

Algirdas Lančinskas[1], Julius Žilinskas[1],
Pascual Fernandez[2], Blas Pelegrin[2]

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

We address a discrete competitive facility location problem for an entering firm with a binary customers choice rule and an asymmetric objective function. A heuristic optimization algorithm which is based on ranking of candidate locations and specially adopted for the discrete facility location problems is designed. The proposed algorithm is experimentally investigated by solving different instances of the facility location problem with an asymmetric objective function.

Dynamic Car Rental Pricing

Karolis Lašas[1,2], Jonas Uus[1,2], Tomas Krilavičius[1,2]

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

Due to high competition in car rental market, dynamic pricing is becoming a necessity for companies that want to gain competitive advantage. Dynamic pricing is commonly used in airplane tickets pricing, it is used for Uber and Bolt services pricing, and it is coming to hotel pricing as well. The rental price is determined by many criteria such as: pick-up and return location, time, duration, seasonality, type of car, number of free cars, time before pick-up and many more. In this research we analyze application of artificial intelligence and optimization for dynamic car pricing. Existing research shows that addition to conventional optimization techniques, machine learning and regression analysis can be used to determine the optimal price, i.e. the maximal price when customers still choose to rent a car. In this study we use linear regression, Light GBM (gradient boosting framework), XGBoost, LSTM, LSTM with CNN and deep reinforcement learning methods to forecast optimal rental car prices using historical data. We compare the effectiveness of chosen methods, report them, and provide future research directions, such as data augmentation (external features, such as events, weather, strikes), a combination of different methods and classical optimization methods.

Data-driven Approaches for Predictive and Preventive Medicine

Tatjana Loncar-Turukalo

Faculty of Technical Sciences
University of Novi Sad, Serbia
turukalo@uns.ac.rs

As the worldwide population grows and the access to health care is increasingly being demanded, the need for a paradigm-shift towards pervasive, predictive and personalized health care decisions becomes evident. Connected health, a technology-enabled model of health management, relies on sensing, communications, and analytics, leveraging that technology to deliver more efficient and cost-effective health care models. The acquisition of health-related data from the patient in a pervasive and seamless manner will allow for the use of powerful analytical tools not only in predictive purposes but as well research-wise to help understand the origins and silent progression of the disease. There are still numerous obstacles to this aim mainly related to slow translation of discoveries in life sciences into more effective therapies, the lack of evidence in the literature on advantages of connected health solutions, and only scattered efforts in interdisciplinary education in healthcare technologies. In this talk, we would address some examples of data-driven approaches supporting medical decision making, from clinical level examples to the research on human microbiome and its association to disease. With an emphasis on challenges associated with different data types, the talk will offer some methodological approaches to tackle these problems, pointing out some important issues from a signal processing and machine learning perspective.

Transformation Algorithms of UML Models Generation Process. UML Dynamic Models Generation Examples

Audrius Lopata[1], Ilona Veitaitė[2]

[1] Kaunas University of Technology
[2] Kaunas Faculty, Vilnius university
ilona.veitaite@knf.vu.lt

Information system (IS) life cycle's design models transformation and generation is a meaningful process in Model Driven Engineering (MDE). Theoretical and practical researches for MDE have extremely advanced recently. It was done in managing the grow of complexity within IS during their development and support processes by increasing the level of abstraction and using different types of models as suitable information storage – as problem domain's knowledge storage. As models increase in use for creating and developing information systems, the possible transformation of these models from different knowledge storages increase in importance. The main scope is to present transformation algorithms of Unified Modelling Language dynamic models generation from Enterprise Model (EM) by theoretical representation of the algorithms and their depiction by examples of generated UML models. The transformation algorithms are described by steps and illustrated by particular UML models examples. Research Action designated as COST Action CA15123: The European research network on types for programming and verification (EUTYPES).

Document Classification to Domains of Use: Lithuanian Case

Justina Mandravickaitė[1,2], Tomas Krilavičius[1]

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

We discuss an experiment on classification of Lithuanian texts according to their domain (area of use), i.e. their functional style. Our experiment includes document classification into 3 functional styles of the Lithuanian language, namely administrative, publicist and scientific styles. We compare classification results of 5 algorithms: LDA, QDA, k-NN, SVM with kernel function and Naïve Bayes and use 8 quantitative linguistic indicators as features (Average Token Length, normalized h-point, vocabulary richness measure R1, normalized Repeat Rate, Moving Average Type-Token Ration, Thematic Concentration, Activity and Verb Distances). The main difficulty in this experiment was the vast diversity of texts, e.g. the administrative style includes documents from diplomatic letters and memorandums to applications and instructions, while the scientific style includes texts such as scientific articles, textbooks as well as articles of popular science, and the publicist style covers news articles, commentaries, essays, reviews, etc. Quantitative linguistic indicators selected as features for distinguishing among the styles were chosen taking into consideration their (more or less) independence of text length as well as qualitative features of separate functional styles discerned by Lithuanian linguists. For the administrative style, SVM was the most effective (96.5% of texts were classified correctly), for the publicist style – LDA (98.9% of texts were classified correctly), and for the scientific style – QDA (93.1% of texts were classified correctly). We achieved the best F-measure score with SVM with kernel function (94.7% – for the administrative style, 98.9% – for the publicist style, and 85.9% – for the scientific style).

Global Optimization in Drug Discovery

Pilar Martínez Ortigosa

Department of Computer Science
University of Almeria, Spain
ortigosa@ual.es

The discovery of new drugs is a very expensive process, frequently taking around 15 years with success rates that are usually very low. New techniques based on principles of Physics and Chemistry have developed about three or four decades ago for the computer simulation (mainly using high-performance computing architectures) of systems of biological relevance. Computational chemistry was later applied for processing large compound databases, and also for predicting their bioactivity or other relevant pharmacologic properties. Using this approach, it was shown that it was possible to use such computational methodology to pre-filter compound databases into much smaller subsets of compounds that could be characterized experimentally. This idea was named Virtual Screening (VS), and it reduces the time needed and expenses involved when working on drug discovery campaigns. Among the most widely used VS approaches, Shape Similarity Methods compare in detail the global shape of a query molecule against a large database of potential drug compounds. Even so, the databases are so enormously large that, in order to save time, the current VS methods are not exhaustive, but they are mainly local optimisers that can easily be entrapped in local optima. It means that they discard promising compounds or yield erroneous signals. In this work, we propose the use of efficient global optimisation techniques, as a way to increase the quality of the provided solutions. In particular, we introduce OptiPharm, which is a parameterizable metaheuristic that improves prediction accuracy and offers greater computational performance than most known VS algorithms. OptiPharm includes mechanisms to balance between exploration and exploitation to quickly identify regions in the search space with high-quality solutions and avoid wasting time in non-promising areas.

Investigation of User Fatigue Based on Input Behavior

Dalius Mažeika, Kšyštof Dunovski

Vilnius Gediminas Technical University
dalius.mazeika@vgtu.lt

Most of the computer users suffer from mental and physical fatigue. A tired person can make mistakes and violate sensitive data integrity or cause other problems in IT systems. Thus it is vital to keep the exhaustion of the user in check.

In this research, we address the problem of identifying user fatigue level through the analysis of input behavior from the mouse and keyboard. Firstly, a specialized tool was developed, which was used to gather data about the keystroke dynamics and mouse motion characteristics. The following data were acquired from the keyboard: up-down, down-down, and holding time of the keys as well as keystroke frequency. Motion speed and hold time of the mouse key were gathered from the mouse. After the tool was created, a static text, as well as a combination of mouse inputs, were given for the volunteered users to make inputs. Corresponding data was gathered and labeled according to the user fatigue level.

Neural network, K-Means as well as classification and regression tree algorithms were used to build user fatigue prediction models. Investigation on different datasets was performed, and the correlation between results obtained from keyboard and mouse datasets was analyzed. Analysis of the resulting accuracies of the models was performed as well, and corresponding conclusions about the capability of predicting user fatigue based on input behavior were made.

Data Analysis Methods from the Perspective of Experimental Sciences

Gerda Ana Melnik

Institute of Data Science and Digital Technologies
Vilnius University
gerda.ana.melnik@gmail.com

Data analysis methods are by and large developed and used by experts, having strong backgrounds in math, computer science and statistics. However, these methods are also essential in the toolkit of the “amateur” data-scientist, working in such important fields as research in medical science, psychology and others. As data science and machine learning get a central role in experimental sciences, researchers in these fields have to face many challenges to keep up with the developing technologies and analysis methods. In certain fields, such as experimental psychology and psycholinguistics, applying complex analysis methods and using sophisticated modelling techniques has almost become a prerequisite for publication in good journals, although such techniques and methods are rarely explicitly taught in laboratories and similar scientific institutions. In this talk we will make an attempt to overview the challenges that psychologists and other specialists in experimental sciences are facing when making decisions about the experimental design of their studies and about the data analysis methods to be (and not to be!) used in their research. Furthermore, we will explore this issue in light of the “replication crisis” – the ongoing methodological crisis demonstrating that many scientific results, especially in life and social sciences, are difficult or impossible to reproduce. The positive impacts of this crisis on promoting better and more rigorous science will be discussed.

On Issues Related to Interval Type-2 Membership Function Development

Jolanta Miliauskaitė[1], Diana Kalibatienė[2]

[1] Institute of Data Science and Digital Technologies
Vilnius University
[2] Information Systems Department
Vilnius Gediminas Technical University
jolanta.miliauskaite@mif.vu.lt

Fuzzy approaches that are proposed to describe uncertain, impressive or vague concepts, are based on the development of membership function (MF), which reflects what is known about the linguistic variables in the application domain. Recently, interval type-2 MF (IT2MF) is becoming widely applicable, since its employment for managing uncertainty (i.e., uncertain meaning, uncertain measurement, and noisy data) allows us to obtain more desirable results. IT2MF can control the blurring better than Type-1 MF (IT1MF) because in IT2MF the Footprint of uncertainty (FOU) provides an extra degree of freedom. However, a non-trivial problem exists in how to construct the most appropriate IT2MF that has the best-fit representation of the analyzed problem. Many authors propose their ways to develop IT2MF using a certain technique in a particular application domain, like video streaming, energy-efficient and load balancing in cloud environment, quality of web service, continuous adaptations, building information granules, pattern recognition, etc. In this research, we aim to systematize the main issues concerning IT2MF development.

Although IT2MF provides us richer information and present a stronger ability to handle uncertainties than IT1MF, its application is much more complex than IT1MF. The main issues are presented as follows. First, this complexity appears through growing computational complexity, especially in the fuzzification and defuzzification processes, i.e. the number of calculations insurmountable because of its combinatorial nature. Second, there is a need to reduce uncertainty by defining IT2MF boundaries that may vary in some range depending upon a large amount of non-deterministic information and noise data in real data. It leads to the fact that T2MF boundaries and shape are not clear in many cases. Third, recent modern technologies produce a huge number of streaming data (i.e., data stream). Consequently, off-line and traditional methods are becoming impractical (i.e., huge memory, long computational time and retraining the developed model). The main conclusion is that there is a need for a general approach for developing IT2MF, which systematizes and generalizes the analyzed approaches into a general methodological framework of developing IT2MF (GMFT2) and proposes a way of solving the mentioned issues.

High Performance Computing Techniques for IMRT Plan Optimization

Juan José Moreno[1], Dmitry Podkopaev[2], Janusz Miroforidis[2], Ernestas Filatovas[3], Ignacy Kaliszewski[2], Ester Martín Garzón[1]

[1] 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

Intensity-Modulated Radiotherapy (IMRT) is a common technique for cancer treatment, allowing to precisely control the geometry and intensity profile of radiation beams. The goal of IMRT planning is to deliver prescribed radiation doses to tumorous tissues, while minimizing the dose inevitably received by the surrounding organs and normal tissue. In the standard approach to IMRT plan optimization, the decision variables represent two-dimensional intensity profiles of several radiation beams aimed from different directions (fluence matrices). The dose deposited in the patient body by a given IMRT plan is modelled by a sparse matrix with thousands of columns (the resolution of the intensity profiles) and hundreds of thousands of rows (the resolution of the model representing the patient body). Thus, a large part of computation resources during optimization is spent on operations with sparse matrices. In this work we explore the effectiveness of a gradient-based optimization method for IMRT plan generation combined with cutting-edge technologies of high performance computing. This way multiple alternative plans can be generated in a limited time span, giving radiologists more freedom in exploration of various planning opportunities.

Lithuanian Speech Synthesis Using Deep Neural Networks

Gediminas Navickas

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

Until now, the best results for Lithuanian speech synthesis were achieved using concatenative unit selection methods. In this work, Lithuanian speech recognition using various neural networks is investigated. The main focus of experiments is on Recurrent Neural Networks (RNN) and their architecture called Long Short Term Memory (LSTM) networks. The main advantage of these networks is the ability to model sequential data, in our case speech data. Feedforward DNNs are compared with LSTM networks and their modification Bidirectional Long Short Term Memory (BLSTM) network. Experiments were performed using Merlin – an open source speech synthesis toolkit which was modified to fit Lithuanian language synthesis.

Reverse Clustering, Classification and Extrapolation: Some Basic Interpretations

Jan W. Owsiński, Jarosław Stańczak,
Sławomir Zadrożny, Janusz Kacprzyk

Systems Research Institute
Polish Academy of Science
jan.owsinski@ibspan.waw.pl

The paper presents the general prerequisites of research on the so-called "reverse clustering" approach against a broader spectrum of issues, primarily related to the question of "classification". We mainly deal with the potential interpretations and uses of the approach, as set in the framework, in which classification is usually considered, our considerations being illustrated by a series of examples.

The gist of the problem of reverse clustering is as follows: we deal with some multidimensional data set *X*, composed of *n* objects or observations, together with its assumed or given partition P_A; for these data, we wish to find the (set of parameters of the) clustering procedure that, when applied to *X*, would produce the partition of this set, denoted P_B, that is as close as possible to the initial given P_A. The set of parameters of the clustering procedure, denoted *Z*, is understood in a truly broad manner, namely encompassing (a) the very choice of the clustering algorithm; (b) the basic parameters of the algorithm (e.g. the number of clusters, the distance threshold, etc.); (c) the distance measure definition; (d) the weighing (ultimately: the choice) of variables. Of course, *Z*, as a vector of "variables", is not uniquely defined in the sense, e.g., that for various clustering algorithms different parameters are accounted for. Further, the space of search is in general very awkward and that is why we decided to use the genetic algorithms to find P_B. Altogether, we minimise some kind of distance between P_A and P_B by appropriately choosing the coordinates of *Z*.

We analyse the relation of the above outlined problem and approach to that of the standard problem of classification, along with its potential various interpretations, and, in this setting, we propose different potential interpretations of the general approach of reverse clustering, presenting also the respective examples of its application for purposes of illustration of these various interpretations.

A New Collection of the Blockchain Platforms

Remigijus Paulavičius, Saulius Grigaitis, Ernestas Filatovas

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

The literature on the application of blockchain technology is extensive and grows at a swift pace. There already exist several collections of blockchain platforms presented in the literature. However, they are limited and focused on particular subclasses. Moreover, as far as we are aware, there is no systematic and comprehensive data library available for the evaluation of the broad class existing and actively developed, as well as newly emerging blockchains platforms.

In this work, we introduce an actively growing online collection of blockchain platforms, BlockLib, gathered from various sources (such as social websites, blogs, wikis, forum posts, source codes, conference proceedings, and journal papers), and devoted to facilitate research on blockchain platforms. The library is designed as an open resource to which other researchers and the blockchain technology community can easily contribute. By doing this, we hope that the blockchain community will help us to fix all errors and inaccuracies, add new data, and keep this collection growing and up-to-date.

Domain Sensitivity Issues in Aspect Based Sentiment Analysis

Mažvydas Petkevičius, Daiva Vitkutė-Adžgauskienė,
Darius Amilevičius

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

While Sentiment Analysis is the task of extraction of subjective information from review texts, Aspect Based Sentiment Analysis (ABSA) is the task of extracting aspect terms and their related sentiment polarity, allowing to monitor the dynamics in positive and negative sentiments expressed towards specific aspects of a product or service for extracting valuable targeted insight. As an instrument for marketers and consumers, ABSA must be aligned with Named Entity Recognition (NER), as in our case to monitor "big picture" of the opinions on university study programmes in social media.

In general, a major challenge of ABSA is the domain sensitivity of both the opinion terms and the aspect terms. In addition, named entity terms are also domain sensitive, especially in social media, where wording is significantly different from normative language, that is the construction of a stable dictionary is practically impossible.

Furthermore, a big challenge is also in linking named entity terms with aspect terms, thus building NER-aspect combinations as opinion targets. Thus, ABSA is the task of co-extracting NER terms (usually in the form of a hierarchical NER-aspect structure), aspect terms and opinion terms, as well as establishing corresponding links between them for a certain domain.

Supervised learning algorithms handle this domain sensitivity challenge well in cases where labeled data from the target domain is provided for training. However, generating labeled data is a labor-intensive and costly effort that requires great human expertise. Corpus data for specific domains of under-resourced languages, such as the Lithuanian language, gives additional challenges. An alternative approach is to apply word embeddings in the initial phase of term extraction, as well as in the production phase, such an approach considered as some kind of lightly-supervised ABSA for under-resourced languages. This approach requires less labeled training data, making its deployment faster and cheaper, and, therefore more practical.

The paper presents the concatenation results of three neural network classifiers (NER, aspect and sentiment), domain sensitive term extraction and a lightly-supervised training approach for a specific domain of social media, covering the use of word-embedding similarities between the acquired terms and sets of generic NER, aspect and sentiment terms as features.

Game Scene Generation by Ventura Creativity Criteria Model

Aurimas Petrovas, Romualdas Baušys

Vilnius Gediminas Technical University
aurimas.petrovas@vgtu.lt

Computational creativity is a growing field, which explores engineering science and philosophy of computational systems, which to the unbiased observer would deem to be creative. As yet, there is no consensus on how to evaluate creative systems as there is a lack of agreed-upon definitions of creativity between psychologists, philosophers and cognitive scientists. There is an argument that computational creativity doesn't need to resemble human creativity. The Ventura model tries to define computational creativity criteria for digital creative content generation. Video games are an interactive media field, which contains a lot of digital content types. There are four main types of creativity evaluation: 1. Person – studies creative agent perks (human), which makes work creative. 2. Process – studies what types of actions are made, that make him creative. 3. Product – a study of creative asset or work, which is considered creative. 4. Press – a study of culture acknowledgement of creative work. Common problems, which are encountered during the computational creativity evaluation process: 1. There is no justification for why certain creativity justification criteria are used. Some criteria limit the creativity of the creative agent. 2. The judgement of non-experts. 3. Bias against computers. 4. Corresponding creative field evaluation criteria. The focus of this research is the implementation and analysis of game scene level design using Ventura philosophy and usage of good level design principles to define indicators for automatic content generation.

Detection of Suspicious Activities for Patient Privacy and Security Management Using Electronic Health Record Access-Log Data

Roma Puronaitė[1,2,3], Rolandas Bėrontas[1],
Danielius Dzekunskas[1,4], Justas Trinkūnas[1,4]

[1] Vilnius University Hospital Santaros Klinikos
[2] Institute of Data Science and Digital Technologies, Vilnius University
[3] Faculty of Medicine, Vilnius University
[4] Vilnius Gediminas Technical University
roma.puronaite@santa.lt

Electronic health records (EHRs) used intensively in the daily workflow of the hospital. Medical staff must have access to medical records for patient care. Data in EHR is sensitive and must be protected due to privacy policies and patient safety. Due to the large amount and complexity of data, auditing user activities has become a more complex and time-consuming task, while a rapid response to security incidents is required. For these reasons, partial automation of the data security process can make the detection and management of suspicious activities more effective. To identify possible suspicious activities, we have analyzed two-week access-logs of the Vilnius University Hospital Santaros Klinikos Electronic Health System. Each record contains time-stamped data along with additional information, such as the reason for access, unit of the patient, user and the role of the user. We used network analysis to identify user behavior, anomalies, and suspicious activities. We have created rules for annotating access-log records, classifying them from one (lowest risk) to five (highest risk). In the final step, we developed a patient privacy and security management tool, in which we implemented an automated solution with model retraining using ML.NET.

Maturity Models for Medical Informatics Management – A Data Analytics Maturity Model

Álvaro Rocha

University of Coimbra, Portugal
amrrocha@gmail.com

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

Player Recognition by Cursor Movement in Computer Game

Petr Romov [1, 2]

[1] GOSU Data Lab
[2] Vilnius Gediminas Technical University
petr.romov@vgtu.lt

The key issue of this work is to understand whether the cursor movement during the use of a computer program can provide sufficient biometric information to recognise the user. Computer games are a good testbed for the experiments because they collect a large amount of accurate data about player behaviour in the form of replays. This data is not sensitive, but still, it is bind to real player's accounts. In this paper, we consider Dota 2 – the popular real-time strategy game. We present a method for learning representation of mouse movements in the form of vectors that are compared using the Euclidean distance. Performance of the proposed method is evaluated on the dataset of 100k game replays and 10k personalities. Two recognition scenarios are considered: verification and identification. The method shows a very high correct verification rate (more than 98%) and relatively good identification performance, enough for practical use.

Applications of AI for 5G Spectrum Sensing

Julius Ruseckas, Gediminas Molis,
Aušra Mackutė-Varoneckienė, Tomas Krilavičius

Baltic Institute of Advanced Technology
julius.ruseckas@bpti.lt

Higher number of wireless devices and communication and forthcoming deployment of 5G networks requires more flexible and efficient spectrum sharing approaches. However, to achieve efficient spectrum sharing between non-cooperating networks a fast spectrum scan is necessary. Power, modulation, frequency and bandwidth have to be quickly estimated to adapt to the environment and cause minimal interference for other network users even when the protocol is not known. We propose to apply convolutional neural networks for multi-carrier signal detection and classification because it can measure all above mentioned parameters from one short data sample. Six multi-carrier signal modulations were generated for the classification and detection tasks. We have measured detection probability and classification accuracy over a wide range of signal-to-noise ratios and have estimated the computational resources needed for the task. Moreover, we have studied the impact of signal augmentation during training phase on classification accuracy when only a part of the signal is available. We show that signal four times shorter than 5G radio subframe can be sufficient for the task.

Research is funded by the project “CERTAIN” No. S-LL-18-9 of the Research Council of Lithuania under Polish-Lithuanian Funding initiative DAINA.

Accuracy of Nonparametric Density Estimation for Univariate Gaussian Mixture Models and Application on Municipal Solid Waste

Tomas Ruzgas, Jurgita Arnastauskaitė

Kaunas University of Technology
tomas.ruzgas@ktu.lt

Many distribution density estimation methods are known in modern data analysis, but practically it is not easy to choose an efficient evaluation procedure if the data distribution density is multimodal, and the sample size is not large. Kernel estimates are found most frequently.

Nevertheless, methods are also popular and often used by Ćwik and Koronacki (1997), Friedman (1987), Hwang (1994), Kooperberg and Stone (1991), etc. In some cases, the accuracy of the assessment significantly increases. According to Ruzgas, Rudzkis, and Kavaliauskas (2006), this is stated, if the observables are clustered at first (treating analysed multimodal density as unimodal density mixture), and the popular nonparametric estimators are applied to each cluster separately. The present work reveals the extended pilot research presented by Ruzgas, Rudzkis, and Kavaliauskas (2006). This study extends the results of a previous study: combines with popular density estimates a completely new one based on the inversion formula; in order to be able to define different types of problems, the very wide set of distributions proposed by Marron and Wand (1992) is used to study densities; they evolve of computer technology has made it possible to carry out more imitative modelling experiments in less time, therefore, 100000 independent samples were generated for each study, which provides a reasonably high level of confidence in the results obtained. The primary purpose of this research is to assess the performance of several density estimators, focusing on benchmark densities that represent different types of problems that can arise for unimodal and multimodal densities. The evaluation errors' dependency in the cases of the complex data structures in univariate Gaussian mixture models (GMM) is analyzed by Monte Carlo method.

Applications of GMM distributions are quite popular and can be found in various scientific fields for examination of relevant problems. After sample clustering in this work, the mixture components were evaluated using: 1) a kernel estimator with the adaptively selected smoothing width, 2) Friedman‘s proposed algorithm based on the projection pursuit, 3) Kooperberg and Stone log-spline estimator based on the approximation of logarithm of distribution density by the sum of cubic B-splines, 4) Fix and Hodges k-nearest neighbor density estimator, and 5) the proposed inversion formula modification for density estimator. A pilot comparative analysis several popular non-parametric estimators’ precision showed that the Friedman procedure was most effective in the majority of examined multivariate cases where the mixture components “clearly” separate, and kernel estimator was more accurate with the small sample size. The application of density estimates is illustrated using municipal solid waste. Data collected in Kaunas (Lithuania) have a similarity pattern in the shape of kurtotic unimodal density. Based on the density estimation, goodness of fit tests have shown that the municipal solid waste fractions densities of Kutaisi (Georgia), Saint-Petersburg (Russia), and Boryspil (Ukraine) are not statistically significant compared to the densities of Kaunas.

Tax Fraud Reduction Using Analytics in an East European Country

Tomas Ruzgas[1], Laura Kižauskienė[2], Jurgita Arnastauskaitė[1,2]

[1] Department of Applied Mathematics, Kaunas University of Technology
[2] Department of Computer Sciences, Kaunas University of Technology
jurgita.arnastauskaite@ktu.lt

Tax administration institutions recently face the challenge of effectively identifying companies that avoid paying taxes. European Union countries are no exception. When dealing with tax evasion problems, tax administrators are confronted with limited resources and often rely on traditional tax audit tools that are time-consuming and labor-intensive. As a result of this established practice, governments are losing a lot of tax revenue. The main objective of this study is to increase the efficiency of the detection of tax evasion by applying data mining methods in East European country (Lithuania) with rapidly developing economics, with respect to affluence-related impacts. Based on data mining methods, various models are developed to address segmentation, risk assessment, behavioral templates and tax crime detection. The results show that the developed data mining technique really makes the detection of tax evasion more effective and is able to extract hidden original knowledge from data that can be further used to efficiently reduce the losses resulting from tax evasion. The methods, software, and findings of this study may assist the experts, decision-makers and scientists performing prediction of tax fraud detection, especially in developing countries.

Application of Vector Fractal Brownian Field Model for Data Extrapolation and Kriging

Leonidas Sakalauskas, Neringa Urbonaitė

Institute of Data Science and Digital Technologies
Vilnius University
neringa.urbonaite@mif.vu.lt

The model of random vector Brownian field (VBF) is developed and numerically implemented for multivariate kriging and data fitting. The aim is to estimate model parameters using multivariate semivariograms and compare that with maximal likelihood estimation. The vector field with exponential covariance function is considered as well, and it is shown, that in limit it tends to considered VBF. The efficiency of models has been tested by applying the computer simulation by Monte Carlo method. Evaluated parameters are used for the kriging model to estimate data values in unknown areas. The prognosis results from each modeling have been produced. The application to emotion recognition for pictures is discussed.

Evaluating the Cause-Effect Relationship for Likert Scale Items

Virgilijus Sakalauskas, Dalia Krikščiūnienė

Vilnius University
virgilijus.sakalauskas@knf.vu.lt

Likert scale items are used for surveys exploring attitudes by collecting responses to particular questions or groups of related statements. The common practice is asking respondents to express their level of agreement by applying the seven or five-point scale from ‘strongly disagree’ to ‘strongly agree’. Although the Likert scale methodology serves as a powerful tool for asking attitudinal questions and getting measurable answers from the respondents, the surveys fail to identify the level of importance of the individual questions used for characterising the explored phenomena. Moreover, the Likert scale methodology does not enable to distinguish which of the questions are the causes, and which of them are the effects of the explored problem. The objective of the research is to propose an original method for evaluating the cause-effect relationship and strength of interdependences among Likert scale items. The modified influential relation map built for Likert scale items revealed improvement scopes by defining causal relationships and significance of the questions of the survey and added value for long-term strategic decision making. The viability of the proposed model is illustrated by a case study of service quality survey data collected at the rehabilitation hospital in Poland.

Connected Vehicle Reference Model for Security Risk Management in Internet of Things

Raimundas Savukynas

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

The Internet of Things (IoT) is a global network of connected devices and processing systems to exchange or accumulate data and information generated by different users of embedded sensors in the physical objects. By identifying, collecting data, processing, and using communication capabilities, the IoT allows the full use of physical objects for various services, ensuring high security and privacy requirements. Among the privacy, environment, energy, and other concerns, security plays an important role, as every physical object connected to the IoT causes significant security threats that are related to the authenticity, confidentiality, integrity of data and services of the secret data of their owners or users. In cases where such data is intercepted and used for non-intended purposes, it may lead to the severe damages of the valuable system and environmental assets. Although the provided IoT benefits are unlimited, many privacy and security challenges are related to the connected vehicle. The connected vehicle uses a network, sensors, and electronic control unit (ECU) to control functions of the vehicle and to connect to other system entities. This way, it exchanges the available information about the car location, current environment, driving direction, condition of the driving, and status information necessary for the connected vehicle device control. Due to connectivity, security vulnerabilities from a single connected vehicle can propagate further and lead to hazards affecting multiple connected vehicles at once. Hence, an attacker can physically change the connected vehicle ECU and provoke wrong driving decisions. The unexpected change of the connected vehicle infrastructure from a passive system to an active system makes it very susceptible to security threats, which leads to safety hazards. In this work, we introduce the connected vehicle reference model for security risk management in IoT, help to discover and explain security vulnerabilities, defining security risks, and introducing security countermeasures.

Creditworthiness Estimation from Aggregated Bank Account Transactions Data

Dovilė Servaitė[1,2], Monika Zdanavičiūtė[1,2],
Rūta Užupytė[1,2], Tomas Krilavičius[1,2]

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

Banks have long been evaluating a person's creditworthiness in terms of client's financial obligations, income, seniority and other data that would not be obtained without special permission. However, credit institutions were able to obtain data provided by customer, while, by using several accounts in different institutions, both customer and bank, were not able to obtain a full picture financial behavior profile. Nowadays, PSD2 directive allows to access all customers' transactions data in different financial institutions via open APIs, in such a way providing means to aggregate all customers accounts data. Moreover, modern artificial intelligence and statistics methods allow in-depth profiling of customer behavior. In these research, clustering and time series prediction methods are applied for customers financial behaviour profiling. We estimate customers creditworthiness using clustering, and use time series analysis to predict money flow in accounts for a selected time horizon. Various estimated monthly balance features of the account are used for clustering, such as: balance average, minimum, maximum, etc. Davies Bouldin, Dunn index and Calinski Harbasz methods are used to determine the number of clusters. Dunn index aim is to identify sets of clusters that are compact, with a small variance between members of the cluster, and well separated, where the means of different clusters are sufficiently far apart. The K-means method is used for clustering. Clusters allow to classify accounts into different credit score categories and determine who qualifies for a loan, to evaluate the potential risk posed by lending money to consumers and to mitigate losses due to bad debt. We calculate monthly account balances as time series. Time series analysis, which investigates whether a time series is stationary, has seasonality or trending. This analysis helps to select the required forecasting model. The customized model shows account's monthly balance for the selected horizon.

Diagnostic Effectiveness of Thermography in the Trauma Presence on the Professional Basketball Players

Augustas Skaudickas[1], Gabija Skaudickaite[2],
Dominykas Meskauskas[2], Vincentas Veikutis[1]

[1] Academy of Medicine Lithuanian University of Health Sciences
[2] Lithuanian University of Health Sciences Gymnasium
vincentas.veikutis@lsmuni.lt

Professional athletes with previous, not fully cured traumas have a 2–3 times greater chance of repeated damage, gradual disability, shortening or complete termination of a sports career. The aim of the work was to determine body temperature distribution contemporizing it with the presence of acute or chronic trauma.

The research was carried out with the athletes of the Kaunas “Žalgiris” basketball team (n-15). Full-body thermoscans were performed before and after the exercise according to the special protocol, giving a tumble to most typical for basketball injuries areas: legs (knees, ankles) and spine. In order to compare the anatomical and thermographic images, we have chosen a scheme for body segmentation. Thermographic images were processed by FLIR TOOLS 2.0 analyze software and the SPSS 20 statistical package.

Thermographic data after training was much less informative because of increased muscle warming masked body temperature anisotropy. We founded as the high-risk area right and especially left ankle parts, 5 athletes (33%) had this trauma. Also, we noticed that the weaker and more traumatized is the left leg (73,3%). This is usually caused by a physiological factor since athletes have it as supportive and the take-off is from the left leg. Incidentally, there was no relation between the severity and localization of the attackers and defenders and their injuries. Serious spine injuries were found on 2, less significant other part injuries – on 80% of players. The sensitivity of the method was 76±21% and the specificity - 82±4% respectively.

Conclusions:

1. The study allows for very precise identification of the traumatic site and the degree of damage. The results are much more accurate done before training.
2. The determined inflammatory expression on individual persons fully reflected objectively sensed complaints of pain or discomfort of movement.
3. Body segmentation and protocol allow dynamically evaluate and dynamically monitor each athlete’s physical status.

Network Configuration Impact on IoT Performance

Julius Skirelis, Dalius Navakauskas

Vilnius Gediminas Technical University
julius.skirelis@vgtu.lt

5th generation networks (5G) enable new possibilities in areas like automated factories, autonomous vehicles, smart cities and many more. Interconnection between smart data sources – Internet-of-Things and data processing devices – has its own limits. The purpose of this research is to reveal what configuration, network topology or network parameters have a greater impact on overall IoT network performance. Three different network topologies with known link parameters were simulated using a network analysis tool, and terms of success and faulting data packets ratio were measured. Additionally, data retransmission between nodes service was simulated, the evaluation of service duration results was performed. Exclusion of static errors in the simulation was performed by introducing controlled jitter multiplier. After a total of 1800 simulations, namely: 2 different network configurations – classic Internet and IoT; 3 different network topologies; 3 different parameter sets with varied bandwidth and delay between the links, with repetition of 100 times for each combination, the analysis of results was performed. The direct side-by-side comparison of two different network configurations results confirms that:

- the packet failure rate on the Internet configuration is more sensitive to the doubled bandwidth or delay;
- the dispersion value depends on the mean ratio value, and doubled delay of the link has the highest impact on dispersion increase;
- doubled bandwidth or delay for the Internet configuration narrows a distribution of the mean ratio value for successfully delivered packets between the nodes;
- non-fully scalable IoT network is more resistant to the network parameters fluctuation and results in better mean packet delivery duration than Internet configuration.

Machine Learning for Data Quality Monitoring and Data Certification in the CMS Experiment at the LHC (CERN)

Mantas Stankevičius

Institute of Data Science and Digital Technologies
Vilnius University
mantas.stankevicius@mif.vu.lt

The Compact Muon Solenoid (CMS) is a general purpose detector working at the CERN Large Hadron Collider (LHC). Physics and detector status data are continuously accumulated with a rate close to 1 kHz, making extremely difficult to perform a real time data monitoring with fine time granularity. Teams of shifters have the duty to check recorded data in order to spot possible anomalies in the detector to avoid using the affected data in Physics analyses. This procedure is referred as Data Certification (DC) and has a primary importance to obtain good Physics results. CMS has a Data Quality Monitoring (DQM) group aimed to accomplish this task efficiently. Machine Learning (ML) aided automated tools can analyse large volumes of data in close to real time and help to detect data quality flaws and failures with lower latency. Efficient anomaly detection with alarm capability helps to save an expensive running time and to collect more valuable physics data. This work summarizes the current state of ML techniques applied to DQM and DC, as well as the on-going efforts trying to greatly reduce the granularity of this process to a few tens of seconds taking advantage of Artificial Intelligence.

Adapting Virtual Environments Based on Patients' Anxiety During Virtual Reality Exposure Therapy

Justas Šalkevičius, Robertas Damaševičius

Kaunas University of Technology
justas.salkevicius@ktu.lt

Nowadays, psychologists have many tools and approaches for the treatment of anxiety disorders, among them are Virtual Reality Exposure Therapy (VRET) systems. In these systems patients are confronted with feared stimuli in controlled virtual environments. However, patients can have different sensitivity for particular stimuli and it can induce different levels of anxiety during VRET sessions. Therefore, it's important to create ways for the psychologist to adapt virtual environment and individualize the treatment for each patient. Changes in patient's anxiety can be observed by measuring bodily reactions controlled by sympathetic part of the autonomic nervous system. Thus, these changes can be tracked through sensors which measure biofeedback signals like galvanic skin response, blood volume pulse and skin temperature. Collected signals then can be processed and used to train the classification model for anxiety detection using machine learning methods. Here we present a cloud based virtual reality exposure therapy system with integrated anxiety recognition framework based on our suggested Virtual Reality as a Service (VRaaS) model. The created system contained virtual environments for the treatment of the social phobia and public speaking anxiety. Additionally, the system provided options for the psychologist to adapt these environments during the VRET sessions based on patient`s anxiety in the real time. Finally, our system was used by psychologists in the clinical experiment and the case study for treatment of anxiety disorders.

Simulation-Based Multi-Objective Optimization Methods for Business Process Optimization

Aleksandr Širaliov, Olegas Vasilecas

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

Business process optimization (BPO) is the focus of all successful business companies. Optimization, itself, is known as the process of finding the best solution from all feasible solutions. Simulation-based BPO is an instrument for detailed analysis of processes and further optimization. Various simulation optimization methods, which is understood as simulation-based optimization, are available. It is not obviously clear which optimization method or group is most applicable for BPO. This paper discusses the simulation optimization methods for BPO. Different simulation optimization approaches have been provided in the related papers, however evolutionary algorithms and in specific genetic algorithms are widely used for BPO. Challenging in BPO becomes apparent when solving problems simultaneously against multiple objectives that conflict to each other. Multi-objective optimization involves optimizing a number of objectives simultaneously and evolutionary algorithms are successfully used to solve related problems. One of the objectives of the paper is to provide sufficient information about simulation optimization and with which methods it is used for BPO. In the field under discussion, it also is a challenge to understand the relation between different terms, such as, Business process optimization, Business process simulation, Multi-objective optimization, Multi-criteria optimization, Simulation optimization, Evolutionary algorithms, Genetic algorithms. Due to large amount of the terms and in some case with very similar wording, it is highly important to use them in proper and precisely way. For that reason, as next objectives of the paper, the explanations of such relations as well as meanings of terms are provided. In our days, market is suggesting some simulation optimization software and it becomes challenging to choose the best fit. The brief comparison of simulation optimization software is also available in the paper. Some ideas how to prepare and run simulation-based multi-objective optimization method for BPO has been presented in the paper. The experiments with BPO, have been conducted with simulation optimization software, will be done and the results will be described in the paper. In the end of the paper, conclusions are listed and what assumptions might be addressed in the future studies. Nevertheless, it is necessary to continue research in the area of the simulation-based BPO to achieve all research objectives and overcome all challenges.

The N-Grams Based Text Similarity Detection Approach Using Self-Organizing Maps and Similarity Measures

Rokas Štrimaitis[1], Olga Kurasova[2], Pavel Stefanovič[1]

[1] Vilnius Gediminas Technical University
[2] Institute of Data Science and Digital Technologies
Vilnius University
rokas.strimaitis@vgtu.lt

The word-level n-grams based approach is proposed to find similarity between texts. The approach is a combination of two separate and independent techniques: self-organizing map (SOM) and text similarity measures. SOM's uniqueness is that the obtained results of data clustering, as well as dimensionality reduction, are presented in a visual form. The four measures have been evaluated: cosine, dice, extended Jaccard's, and overlap. First of all, texts have to be converted to numerical expression. For that purpose, the text has been split into the word-level n-grams and after that, the bag of n-grams has been created. The n-grams' frequencies are calculated and the frequency matrix of dataset is formed. Various filters are used to create a bag of n-grams: stemming algorithms, number and punctuation removers, stop words, etc. All experimental investigation has been made using a corpus of plagiarized short answers dataset.

The Impact of Data Augmentation Based on White Noise for Noise-Robust CNN-Based Speech Recognition

Gintautas Tamulevičius, Gražina Korvel,
Jolita Bernatavičienė, Povilas Treigys

Institute of Data Science and Digital Technologies
Vilnius University
gintautas.tamulevicius@mif.vu.lt

The languages of small nations or dialects are considered as low-resource languages. It is a big problem if we want to explore data-driven approaches like Deep Neural Networks (DNNs). Deep learning requires a large training set to achieve a correct assessment of model performance. In order to provide a sufficient amount of data, the augmentation procedure should be done.

In the deep learning approach, images are most commonly used as network input data. In the case of speech, the recordings can be converted to the two-dimensional feature space and exported as grayscale images. The traditional image augmentation methods are not suitable for speech signals as they can lead to loss or distortion of key characteristics (for example, time-scale properties). In signal processing domain, researchers apply deformations directly to the acoustic signal before converting it into images. The literature review shows that the main focus of most of the scientific papers covering data augmentation is to increase network performance. In this work, we look at the data augmentation from another perspective. We are seeking for augmentation that can also be used for improving noise robustness of the speech recognition. In this research, a thorough analysis of speech signals in the presence of noise is done. For this purpose, we added white noise to the speech signals before converting it to a two-dimensional feature space. The following signal-to-noise ratio (SNR) levels were considered: 30 dB, 25 dB, 20 dB, 15 dB, 10 dB, 5 dB, and 0 dB. Also, the experiment was extended by testing clear recordings with an SNR value of more than 100 dB. These signals were obtained by applying the Wiener filter.

An experiment was performed on the Lithuanian speech recordings using 111-word utterances by 36 females and 26 males. Different tests were conducted on the same recordings, but in the presence of noise. For the experiment, we chose the time-frequency domain representation called spectrograms.

Eye Blood Vessel Segmentation Using Convolutional Neural Networks

Ričardas Toliušis, Olga Kurasova, Jolita Bernatavičienė

Institute of Data Science and Digital Technologies
Vilnius University
ricardas.toliusis@mif.vu.lt

Eye blood vessel segmentation is an actual problem in biomedical image analysis, since analysis of vessels is crucial for diagnosis of various diseases, such as glaucoma, hypertension, diabetic retinopathy, macular degeneration, etc. Automatic segmentation can support in performing this task but still is challenging due advanced disease lesions, image quality and other causes. Various methods are developed for blood vessel segmentation, but methods based on convolutional neural networks have become most popular. The aim of this work is to develop a new method based on a convolutional neural network for eye blood segmentation.

Evaluation of Fractal Dimension for Some Financial Time Series

Lukas Vabalas

Institute of Applied Mathematics
Vilnius University
lukasvabalas@gmail.com

Fractal dimension can be used as a measure of smoothness of a function, to evaluate the Hurst exponent and thus explore market memory using financial time series. It can also be used to evaluate the number of independent variables to model a dynamic system which generates financial time series. In this work, we used box-counting, Hall-Wood, rodogram, madogram, variogram, and, more recent, FD4 fractal dimension estimators. A Monte Carlo simulation experiment was carried out to compare fractal dimension estimators, with the madogram estimator performing the best in case of this experiment plan. We analyzed currency exchange rate and gold price time series data sets with 2^{20} observations. The Hurst exponent was evaluated for these time series using different fractal dimension estimators. Further, correlation dimension of time-delay embedding reconstruction of an attractor was evaluated. We show that more observations are needed to evaluate the correlation dimension in higher embedding dimensions to obtain saturated values.

Distortion-Based Audio Augmentation for Continuous Speech Recognition

Jūratė Vaičiulytė, Gintautas Tamulevičius

Institute of Data Science and Digital Technologies
Vilnius University
jurate.vaiciulyte@mif.vu.lt

The amount of training data is an important issue in continuous speech recognition, both neural networks and statistical approaches. Insufficient or inconsistent training data set (which is the case for low-resource languages) may give low quality or over-trained models resulting in poor speech recognition. Nowadays, the principle of artificially increasing the amount of training is widely applied. The amount of data is increased by adding extra noise, modifying time scale, perturbating the speaker's vocal tract length, distorting acoustics features, or applying speech synthesis to obtain additional data.

In this study, we have explored the influence of distortion-based speech data augmentation on continuous speech recognition rate and its robustness. For this purpose, we have employed additive and convolutional noises for the speech recordings (the speech corpus of ~90 hrs was exploited). The white noise was added at various levels to form an additive noise-based subset of the training data. The convolutional noise was imitated with the help of various room impulse responses (we have used the BUT Speech@FIT Reverb Database for this purpose). The amount of distorted data in the training set was formed by randomly selecting utterances and consistently increasing their amount, thus analysing their impact on recognition accuracy and robustness.

The effect of augmented training data on speech recognition was compared to a corpus of ~200 hrs, advantages and disadvantages of each case were analysed.

Learning Phishing Websites URLs Using Long Short-Term Memory Network and Gated Recurrent Units

Paulius Vaitkevičius, Virginijus Marcinkevičius

Institute of Data Science and Digital Technologies
Vilnius University
paulius.vaitkevicius@mif.vu.lt

Phishing remains a continual security threat, creating global losses exceeding 2.7 billion USD in 2018, according to the FBI's Internet Crime Complaint Center. In 2018, the Anti-Phishing Working Group reported 785,920 unique phishing websites detected, with a 69.5% increase during the last five years. Different generations of phishing websites' URLs detection methods have been proposed by the scientific community. Most primitive methods included blacklisting of phishing websites' URLs in the centralised database to be later used by Internet browsers. More recent methods included classical classification algorithms on phishing datasets with predefined features, extracted from phishing websites' URLs, or with automatic feature extraction. Last few years have shown scientific community's attempts to solve phishing websites detection problem using deep neural networks. These methods do not require manual feature extraction since they directly learn a representation from the sequence of characters in the URL. The purpose of this research is to implement algorithms based on Long Short-Term Memory network (LSTM) and Gated Recurrent Unit (GRU), using natural language processing techniques, and to compare the performance of these algorithms with classic classification methods for detection of phishing websites' URLs.

Detecting Maritime Traffic Anomalies with Long-Short Term Memory Recurrent Neural Network

Julius Venskus[1,2], Povilas Treigys[1],
Jolita Bernatavičienė[1], Jurgita Markevičiūtė[3]

[1] Institute of Data Science and Digital Technologies, Vilnius University
[2] Klaipėda University
[3] Institute of Applied Mathematics, Vilnius University
julius.venskus@gmail.com

Analysis and monitoring of Maritime domain awareness is a very important field of human activity. It helps to secure areas such as Oil Rigs, State borders, harbors, sea vessels, sea lanes, wind pow plants, and other offshore structures. Marine vessel traffic intensity rises every year. A huge amount of big data is generated. To monitor and analyze such a quantity of data, human cognitive abilities are insufficient. On the other hand, traditional machine learning algorithms are not capable or unpractical to be used in this type of application. To solve such complex task, other approaches must be researched. In especial, deep neural networks can be considered as an alternative practical approach. In this research, a multistacked LSTM deep neural network is used for abnormal maritime vessel traffic detection. AIS data is prepared to form marine vessel movement multivariate multi-step time series sequences. The prepared sequences are fed to the LSM network and the marine vessel normal traffic model is trained. The error distribution with the covariance matrix is calculated to detect abnormal marine traffic against the normal model learned with the LSM network. The approach is tested with Denmark offshore vessel traffic big data. The results of experiments show the effectiveness of the proposed approach in comprehensive real-world data and should be investigated further for abnormal marine traffic detection.

Development of an SME-oriented Information Security Risk Analysis Expert System Based on Automatically Formatted Knowledge Base

Donatas Vitkus, Vitalijus Gurčinas,
Žilvinas Steckevičius, Nikolaj Goranin

Vilnius Gediminas Technical University
d.vitkus@vgtu.lt

Today almost all companies process data and automate their processes using information technologies. Even the smallest company possesses information and an information system that have to be secured. Information security risk analysis is a compulsory requirement both from the side of regulating documents and information security management decision making process. However, it is a challenging task for small and medium-sized enterprises (SME) because of lack of competence and limited resources. In our work, we present a model of an expert system, dedicated for information security risk analysis for SMEs, which gets a knowledge base automatically from existing information security sources since creation of a knowledge base by expert interviews is timely and expensive. For knowledge base formation, we have used an information security standard ontology. The generated rules were integrated into the JESS-based prototype risk analysis ES, adopted for SMEs. The knowledge base included rules for identification of appropriate assets, calculating impact and probabilities based on the environment of a specific company (infrastructure, maturity level, environment, sector, etc.), while the rules generated automatically from the ontology mainly included information on appropriate security controls. ES questions were adapted for five different SME business profiles types: e-commerce, software development, design, transportation and private medicine clinics. Main factors that determined what questions the system activates were the enterprise size and business profile. If there were facts in the Dynamic facts base, they were loaded into the memory with higher priority than knowledge base facts. After the risk analysis process, the user has to

decide how to reduce the risks. Therefore, ES not only evaluates the risks but also gives some recommendations based on the acceptable risk level. ES was verified with the combined knowledge base (automatically generated and prepared rules). Tests included a comparative analysis of results obtained by the ES and independent human experts that got the same company profile and questionnaire as the ES. The obtained results have demonstrated the suitability of the proposed approach both from the ES use for risk analysis and automatic formation of knowledge base perspective.

Verification of the Method for Multi-Objective Business Process Resource Optimization

Tadas Vysockis, Olegas Vasilecas

Vilnius Gediminas Technical University
tadas.vysockis@vgtu.lt

In the modern business world, there is a frequent need for organizations to improve the structure and usage of resources of their business processes in order to be able compete in today's business environment. This process is better known as optimization. Business process optimization is one of the main activities in organizations, but it is long and usually time-consuming work. One way to optimize processes is to use business process simulation, which allows analyzing various business scenarios under different circumstances, provides an understanding of the most important factors affecting the process. In this context, multi-objectivity is expressed in terms of main business process factors (for example cost, duration etc.). In order to automate this optimization process, it is possible to create a tool, which would help to analyze different simulation results and find the optimal resources set which improves most important process performance indexes.

Therefore, the goal of this research is to develop a method by combining the multi-objective optimization algorithms with business process simulation. This method may result in new approaches and more efficient ways for improving business processes. In this paper, we focus more on the validation of the suggested method part and current results. In addition, we present results of ongoing research on business process optimization based on the multi-objective optimization algorithms. We use these algorithms to find best resources sets for specified fitness function.

The paper highlights main research problems and describes an approach for multi-objective business process optimization. Based on the proposed approach, the optimization prototype was developed and validated. During the validation step we compared our results with the similar results obtained with the Petri Nets based method to assure the reliability of the method. This paper reflects the state of the art in the business process simulation and multi-objective optimization, describes business process multi-objective optimization method. In addition, paper presents validation of the proposed method, obtained results and discussions.

For notes

For notes

11th International Workshop

DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS

Compiler **Jolita Bernatavičienė**
Prepared for press and published by
Vilnius University
Institute of Data Science and Digital Technologies
4 Akademijos St., LT-08412 Vilnius
Vilnius University Press
Saulėtekio av. 9, III bld., LT-10222 Vilnius
info@leidykla.vu.lt, www.leidykla.vu.lt

General sponsors

Research Council of Lithuania

algoritmų sistemos

Main sponsors

NOVIAN

WESTERN UNION WU

VTeX

Sponsors

baltic amadeus

NetCode

NRD Cyber Security

VITP Visorių informacinių technologijų parkas