Area: Analytics, Data Science and Data Mining
Stream: European Working Group: Data Science Meets Optimization

----

Session 132: MA-09: The Role of Mathematical Optimization in Data Science I
Chair(s): Vanesa Guerrero

-----

* 2219
  Tree-structured data clustering
  Derya Dinler, Mustafa Kemal Tural, Nur Evin Ozdemirel

Traditional clustering techniques deal with point data. However, improving measurement capabilities and the need for deeper analyses result in collecting more complex datasets. In this study, we consider a clustering problem in which the data objects are rooted trees with unweighted or weighted edges. Such tree clustering problems arise in many areas such as biology, neuroscience and computer or social networks. For the solution of the problem, we propose a k-means based algorithm which starts with initial centroid trees and repeats assignment and centroid update steps until convergence. In the assignment step, each data object is assigned to the most similar centroid determined by the Vertex Edge Overlap (VEO). VEO is based on the idea that if two trees share many vertices and edges, then they are similar. In the update step, each centroid is updated by considering the data objects assigned to it. For the unweighted edges case, we propose a Nonlinear Integer Programming formulation to find the centroid of a given cluster which is the tree maximizing the sum of VEOs between trees in the cluster and the centroid itself. We solve the formulation to optimality with a heuristic. For the weighted edges case, we provide a Nonlinear Programming formulation for which we have a heuristic not guaranteeing optimality. We experiment with randomly generated datasets and compare the results with those of the traditional k-modes and k-means algorithms. Preliminary results are promising.

-----

* 1507
  Dealing with heterogeneous data via constrained lasso
  M. Remedios Sillero-Denamiel, Rafael Blanquero, Emilio CARRIZOSA, Pepa Ramírez-Cobo

The Lasso has become a benchmark data analysis procedure, and it has been studied in depth and extended by many authors. In particular, we can find in the recent literature some Lasso variants which deal with data collected from distinct strata, as it is standard in many biomedical contexts. In this work we propose a novel variant of the Lasso in which the degrees of reliability of the different data sources are taken into account. However, this is not enough to guarantee the good performance across all the considered sources. Therefore, we shall additionally demand that overall performance measures, as well as performance measures for the data from each source, attain certain threshold values. As a result, a generalized regression model is obtained, addressed by solving a nonlinear optimization problem.

-----

* 3088
  DCA for robust principal component analysis
  Hoai Minh Le, Vo Xuan Thanh

In many real-world applications such as image processing, genomic analysis, etc we encounter very high-dimensional data. Analysis of such data is challenging due to the curse of dimensionality. Principal Component Analysis (PCA) is one of the most popular dimensionality reduction methods to handle high dimensional data. In this work, we deal with Robust PCA (RPCA), a modification of PCA which works well with corrupted observations. RPCA has been succesfully applied to many real life important applications such as Video Surveillance, Face Recognition, etc. RPCA consists in decomposing a given data matrix A as a sum of a low rank matrix and Y and is a sparse noise matrix Y. Thus, RPCA can be formulated as a minimization of the trade-off between the zero-norm of Y and the rank of X. It is well-known that the minimization of a zero-norm is NP-hard problem. We approximate the zero-norm by a DC (Difference of Convex function) function. The resulting problem is then a DC program. We investaged a DCA based method to solve the latter. Numerical experiments on several synthetic and real datasets show the efficiency of our DCA-based algorithm.

-----

* 1930
  On mathematical optimization to interpret factor analysis
  Vanesa Guerrero, Emilio CARRIZOSA, Dolores Romero Morales, Albert Satorra

A natural approach to interpret the latent variables arising in Factor Analysis consists of measuring explanatory variables over the same samples and assign (groups of) them to the factors. Therefore, each factor is explained by means of their assigned explanatory variables yielding a straightforward way to give meaning to the factors. Whereas such assignment is usually done by the user based on his/her expertise, we propose an optimization-based procedure which seeks the best transformation of the factors that yields the best assignment between them and given (groups of) explanatory variables.

Session 133: MB-09: The Role of Mathematical Optimization in Data Science II
Chair(s): Vanesa Guerrero

-----

* 2530

* 2529
  On the construction of optimal randomized classification trees
  Cristina Molero-Río, Rafael Blanquero, Emilio CARRIZOSA, Dolores Romero Morales

Random Forests are a powerful prediction tool obtained by bagging decision trees. Classic decision trees are defined by a set of orthogonal cuts, i.e., the branching rules are of the form variable X not lower than threshold c. The variables and thresholds are obtained by a greedy procedure. The use of a greedy strategy yields low computational cost, but may lead to myopic decisions. Although oblique cuts, with at least two variables, have also been proposed, they involve cumbersome algorithms to identify each cut of the tree. The latest advances in Optimization techniques have motivated further research on procedures to build optimal classification trees, with either orthogonal or oblique cuts. Mixed-Integer Optimization models have been recently proposed to tackle this problem. Although the results of such optimal classification trees are encouraging, the use of integer decision variables leads to hard optimization problems. In this talk, we propose to build optimal classification trees by solving nonlinear continuous optimization problems, thus avoiding the difficulties associated with integer decision variables. This is achieved by including a cumulative density function that will indicate the path to be followed inside the tree. Numerical results show the usefulness of this approach: using one single tree, we obtain better accuracies than classification trees and close to Random Forests, being much more flexible since class performance constraints can be easily included.

-----

* 2873
  Stochastic DCA for group feature selection in multiclass classification
  Bach Tran, Hoai An Le Thi, Hoai Minh Le, Duy Nhat Phan

In this work, we consider the problem of minimizing a large sum of DC functions, which is frequently encountered in machine learning. This problem has a double difficulties due to both large-number of component functions and non-convexity of objective function. We propose an efficient Stochastic DCA (DC Algorithm) to solve it. As an application, we consider the feature selection for multiclass logistic regression problem. Numerical experiments on several datasets demonstrate the effectiveness of our method in terms of three criteria: classification accuracy, sparsity of solution and computing time.

-----

* 1473
  Class probability estimation for SVM with asymmetric costs
  Sandra Benítez-Peña, Rafael Blanquero, Emilio CARRIZOSA, Pepa Ramírez-Cobo

Support Vector Machine (SVM) is a powerful tool to solve binary classification problems. However, SVM does not provide probabilities as other classifiers do in a natural way. Many attempts have been carried out in order to obtain those values, as in Sollich P. (2002) and Platt J. et al (1999). Here, a bootstrap-based method yielding class probabilities and confidence intervals is proposed for a novel version of the SVM, namely, the cost-sensitive SVM in which misclassification costs are considered by incorporating performance constraints in the problem formulation. This is important in many contexts as creditscoring and fraud detection where misclassification costs may be different in different classes. In particular, our target is to seek the hyperplane with maximal margin yielding misclassification rates below given threshold values.

-----

* 1253
  Some improvements of SVM and SVR for functional data
  M. Asuncion Jimenez-Cordero, Rafael Blanquero, Emilio CARRIZOSA, Belen Martin-Barragan

Functional Data Analysis (FDA) is devoted to the study of data which are functions and Support Vector Machines (SVM) is a benchmark tool for classification, in particular of functional data. SVM is frequently used with a gaussian kernel, which has a bandwidth, usually considered as a scalar value. With the aim of preserving the advantages of the scalar bandwidth and, at the same time, exploiting the functional nature of the data, in this talk we propose a new functional kernel that is able to weight different ranges of the functional data according to their classification ability. Moreover, the selection of a subset of variables, e.g. time instants, which contain the most relevant information of the data in order to obtain a good classification is also described.

The search of the most informative time points can be easily extended to the functional regression field by means of the well-known technique Support Vector Regression (SVR).

Tuning the associated functional parameters is a challenging task that we express as a continuous optimization problem, solved by means of a heuristic.

Our experiments with synthetic and benchmark real data sets show the advantages of using functional parameters and the effectiveness of our approach.

Session 136: MC-09: The Role of Mathematical Optimization in Data Science III
Chair(s): Philipp Baumann

-----

* 3039
  An optimization model for credit score based offer management in telecommunication sector
  AYSE SEDA YAVUZ, Erinc Albey

In the telecommunication business, companies not only provide voice/text/data services but also utilize some up-sell strategies such as selling mobile phones to customers. These activities maximize customer life time value and minimize risk of churn through contractual agreements covering long time intervals. In this work we study an optimization problem that deals with deciding the most suitable device and contract setting for each customer while considering risk scores. We first consider risk scores calculation methods based on utilizing several customer specific features such as demographic information, voice/data usage history, average invoice amount, payment history, equipment usage routine. After risk score calculation, credit limits are determined, which are further used to assist on the device assignment decision. An assignment model aiming to maximize sales while maintaining the credit amount allocated to customers under certain threshold levels is constructed using mixed integer programming. The resulting problem is then solved using commercial solvers, and for large instances, heuristic methods are applied. The performance of the solution approach is analyzed under several experimental settings.

-----

* 2395
  A matheuristic for binary classification of data sets using hyperboxes
  Derya Akbulut, Cem Iyigün, Nur Evin Ozdemirel

In this study, an optimization approach is proposed for the binary classification problem. A Mixed Integer Programming (MIP) model formulation is used to construct hyperboxes as classifiers, minimizing the number of misclassified and unclassified samples as well as overlapping of hyperboxes. The hyperboxes are determined by some lower and upper bounds on the feature values, and overlapping of these hyperboxes is allowed to keep a balance between misclassification and overfitting. A matheuristic, namely Iterative Classification procedure for Binary classes (ICB) is developed based on the MIP formulation. In each iteration of the ICB algorithm, a fixed number of hyperboxes are generated using the MIP model, and then a trimming algorithm is used to adjust the hyperboxes in a way to eliminate the misclassified samples. Some trimmed hyperboxes and sample assignments are then fixed, reducing the unclassified sample size left for the next iteration. ICB controls the number of hyperboxes in a greedy manner, but provides an overall hyperbox configuration with no misclassification by the end of the training phase. For the test phase, distance-based heuristic algorithms are also developed to classify the uncovered and overlap samples that are not classified by the hyperboxes.

-----

* 2212
  Particle swarm optimization based regression
  Donghee Yoon, SeJoon Park, Nagyoon Song, Suemin Kim, Dohyun (Norman) Kim

Regression analysis is used for predictive and descriptive purposes in various fields. For the descriptive purpose, it very important to obtain an explicit regression function for an output variable. However, many existing regression methods including support vector regression have a difficulty in describing an explicit function that expresses the nonlinear relationship between an output variable and input variables. To resolve this problem, we propose a nonlinear regression algorithm using particle swarm optimization by which the output variable can be explicitly expressed as a function of the input variables. The experimental results show that the proposed method, even with an explicit regression function, performs slightly better than the existing methods regardless of the datasets, implying that it can be used as a useful alternative when obtaining the explicit function description of the output variable using input variables.

-----

* 2253
  Scaling up similarity-based machine learning models: new geometric algorithms for sparse computation
  Philipp Baumann, Dorit Hochbaum, Quico Spaen

Many leading machine learning models, including mathematical programming-based models, require as input similarities between objects in a data set. Since the number of pairwise similarities grows quadratically with the size of the data set, it is computationally prohibitive to compute all pairwise similarities for large-scale data sets. The recently introduced methodology of &quot;sparse computation&quot; resolves this issue by computing only the relevant similarities instead of all pairwise similarities. To identify the relevant similarities, sparse computation efficiently projects the data onto a low-dimensional space that is partitioned into grid blocks. A similarity is considered relevant, if the corresponding objects fall in the same or adjacent grid blocks. This guarantees that all pairs of objects that are within a specified L-infinity distance with respect to the low-dimensional space are identified as well as some pairs that are within twice this distance. For very large data sets, sparse computation can have high running time due to the enumeration of pairs of adjacent blocks. We propose here new geometric algorithms that eliminate the need to enumerate adjacent blocks. Our empirical results on data sets with up to 10 million objects show that the new algorithms

achieve a significant reduction in running time.

Session 616: MD-09:   The Role of Mathematical Optimization in Data Science IV
Chair(s): Vanesa Guerrero

-----

* 3561
  A branch-and-price approach to find optimal decision trees
  Murat Firat, Guillaume Crognier, Adriana F. Gabor, Yingqian Zhang

In Artificial Intelligence (AI) field, decision trees have gained certain importance due to their effectiveness in solving classification and regression problems. Recently, in the literature we see finding optimal decision trees are formulated as Mixed Integer Linear Programming (MILP) models. This elegant way enables us to construct optimal decision trees with different error concerns. In our work, we find the optimal decision trees by employing the Column Generation (CG) approach. To do so, we first reformulated the previously proposed MILP model as a master MILP model. This master model is first reformulated the previously proposed MILP model as a master MILP model. This master model is basically a partitioning problem of the rows in the given data set such that every partition, or so-called segment, of the row set is assigned to a leaf of the decision tree. We obtain the integer solutions by applying a Branch-and-Price search. Our approach of constructing decision trees is successfully tested in real-world benchmark datasets.

-----

* 1124
  Predicting review helpfulness: an open problem?
  Anne-Sophie Hoffait, Ashwin Ittoo

Online customer reviews represent one of the most popular and accessible source of product/service information. E-commerce platforms enable users to vote for review for their helpfulness, which act as indicator of the review’s reliability for other readers. While numerous scientific publications have focused on the topic of predicting review helpfulness, several questions are yet to be addressed. Moreover, the current literature is highly heterogeneous, leading to inconsistent and contradictory results. Our aim with this study is to synthesize and critically assess the state of the art in research on what makes a review helpful and on predicting review helpfulness. Our primary findings reveal the use of highly varying datasets; a huge plethora of distinct features, including some which are counter-intuitive, as the count of n-letters words or of line breaks; the lack of benchmarks for comparing and assessing algorithms’ performance in predicting review helpfulness or the application of machine learning techniques overlooking the statistical characteristics of the data resulting in flawed results. We propose several research directions to overcome these gaps and advance the state of the art, such as a standard features set and algorithms to be used as benchmark for assessing future research. We also propose new approaches based on recent innovations in argumentation mining and deep learning as well as more advanced statistic techniques, such as lasso/ridge regression.

-----

* 2935
  Interpretable deep learning models for financial applications
  Pavankumar Murali

Statistical learning has been used in the past to build several predictive models for financial applications such as risk estimation, attrition prediction, product recommendation etc. A majority of these models are tailor-built for specific datasets and use-cases. As such, they cannot be used easily across use-cases without heavy-lifting to rebuild models. Secondly, the training methods employed do not allow true isolation of feature effects on the outcome. In popular techniques such as logistic regression, one can only decipher the relative importance of a feature in the presence of other features. In this talk, we present a new approach based on deep-learning to address this drawback of traditional machine learning models, while enhancing the capability to extract individual customer-level rationales behind a predicted outcome. We use a modeling approach called attention models, that is inspired by the flexibility and interpretability of generalized additive models. This modeling framework enables us to isolate the training and contribution of each subset of features. For instance, our approach can be used to identify the importance of customer profile-based features vs. that of daily account transactions to the predicted outcome. We present results based on several publicly available datasets to show how attention models perform better than traditional methods, and can provide superior interpretability.

-----

* 595
  Optimized re-ranking approaches for a doctor recommender system
  Hongxun Jiang

Big data nourishes smarter services while detailed consideration promotes satisfying ones. State-of-the-art literature of physician recommenders has put major efforts on accuracy, which used to overlook an emerging phenomenon. A few “superstar” physicians dominate others in the number of times recommended that brings to a quite long waiting time making users impatience, and causes themselves exhausted as well. Diversity does help but cannot eliminate overworks. A novel metric in this paper called workload caps the times a physician recommended to improve the quality of recommendations, based on which a

personalized physician recommender system was developed. The key contributions was an optimized re-ranking approach after supervised learning. The optimization presented as a mix integer programming with subject to the workload constraints for anyone recommended, as well as objective to find the maximum accuracy. It is a large-scale integer programming and has the computation-consuming problem. We supposed a greedy heuristic to find near-optimal solutions that assures the workload standard first and finds alternatives as good as possible. The experimental results proved that the re-ranking approach improved remarkably the quality of recommendations comparing with the benchmark model. In addition, the greedy-based algorithm outperformances the CPLEX-based one in the computational time in all circumstances, while the similar scores in the accuracy of recommendations.

Session 161: TA-09: The Role of Mathematical Optimization in Data Science V
Chair(s): Adam Elmachtoub

-----

* 662
  Small-data, large-scale linear optimziation
  Vishal Gupta, Paat Rusmevichientong

Optimization applications often depend upon a huge number of uncertain parameters with imprecise estimates. We term this setting – where the number of uncertainties is large, but all estimates have fixed and low precision – the “small-data, large-scale regime.” For linear programs with uncertain objective coefficients, we prove that traditional methods like sample average approximation, data-driven robust optimization, regularization, and ``estimate-then-optimize” policies can perform poorly in the small-data, large-scale regime. We propose a novel framework that, given a policy class, identifies an asymptotically best-in-class policy, where the asymptotics hold as the number of uncertain parameters grows large, but the amount of data per uncertainty (and hence the estimate’s precision) remains fixed. We apply our approach to two policy classes for this problem: an empirical Bayes class inspired by statistics and a regularization class inspired by optimization and machine learning. In both cases, the sub-optimality gap between our proposed method and the best-in-class policy decays exponentially fast in the number of uncertain parameters, even for a fixed amount of data. We also show that in the usual large-sample regime our policies are comparable to sample average approximation. Thus, our policies retain the strong large-sample performance of traditional methods, but also enjoy provably strong performance in the small-data, large-scale regime.

-----

* 1068
  Condition number analysis of logistic regression and its implications for standard first-order solution methods
  Paul Grigas, Robert M. Freund, Rahul Mazumder

The elementary probabilistic model underlying logistic regression implies that it is most natural to consider logistic regression when the data is not (linearly) separable. Building on this basic intuition, we introduce a pair of condition numbers that measure the degree of non-separability or separability of a given dataset in the setting of binary classification. When the sample data is not separable, we show that the degree of non-separability naturally enters the analysis and informs the properties and convergence guarantees of two standard first-order methods, namely steepest descent (for any given norm) and stochastic gradient descent. When the sample data is separable -- in which case many properties of logistic regression essentially break down -- the degree of separability can be used to show rather surprisingly that these two standard first order methods deliver approximate-maximum-margin solutions with associated computational guarantees as well. Last of all, in order to further enhance our understanding of the computational properties of several methods, we exploit recent new results on self-concordant-like properties of logistic regression due to Bach.

-----

* 3305
  Discovering optimal policies: a machine learning approach to model analysis
  Fernanda Bravo, Yaron Shaposhnik

We study a novel application of machine learning methods for discovering structural properties of optimal policies in numerically obtained solutions to optimization problems. Our proposed framework provides a systematic approach, which complements numerical experimentation and theoretical analysis, to characterize optimal policies for optimization problems that commonly arise in the context of operations management. As a proof of concept, we apply our framework to core optimization problems, such as inventory replenishment, admission control in queueing systems, and multi-armed bandit (MAB) problems. We demonstrate how to apply standard statistical learning methods to characterize optimal threshold-based policies for the inventory replenishment and admission control problems. For the MAB problem, we derive a new class of policies that generalizes index policies, and develop a new efficient algorithm for computing an optimal index policy. The main contribution of our work is methodological, in proposing and demonstrating the effectiveness of a new machine learning-based method for analyzing optimization problems.
Applying our method leads to new insights about optimal solutions to MAB problems.

-----

* 123

  Smart predict, then optimize
  Adam Elmachtoub, Paul Grigas

We consider a class of optimization problems where the objective function is not explicitly provided, but contextual information can be used to predict the objective based on historical data. A traditional approach would be to simply predict the objective based on minimizing prediction error, and then solve the corresponding optimization problem. Instead, we propose a prediction framework that leverages the structure of the optimization problem that will be solved given the prediction. We provide theoretical, algorithmic, and computational results to show the validity and practicality of our framework.

Session 149: TB-09: The Role of Data Science in Optimization
Chair(s): Patrick De Causmaecker

-----

* 1989
  A data-driven model for routing mobile medical facilities
  Sibel Salman, Burcin Bozkaya, Eda Yücel, Cemre Gokalp

We study a problem related to the delivery of mobile medical services and in particular, the optimization of the joint stop location selection and routing of the mobile vehicles over a repetitive schedule consisting of multiple days. We consider the problem from the perspective of the mobile service provider company, and we aim to provide the most revenue to the company by maximizing the reach of potential customers to the provided services. The problem is a variant of the team orienteering problem with an objective that accounts for full and partial coverage. We utilize transactional (big) data that originate from the customer banking activities of a major bank in Turkey. We analyze this dataset to determine the priority of areas (districts of Istanbul) where services are to be delivered for maximum coverage, coverage parameters and the potential locations of delivery. We formulate a mixed for maximum coverage, coverage parameters and the potential locations of delivery. We formulate a mixed integer-linear programming model and solve it to optimality using CPLEX in reasonable time. Our results indicate an important trade-off between the total number of days of service along with the number of vehicles used, and service coverage level. We compare the results of several models differing in their coverage functions and demonstrate the efficacy of our model.

-----

* 3691
  Constraint learning using tensor (COUNT)
  Mohit Kumar, Luc De Raedt, Stefano Teso

Many problems in operations research require that constraints be specified in the model. Determining the right constraints is a hard and laborsome task. We propose an approach to automate this process using artificial intelligence and machine learning principles. So far there has been only little work on learning constraints within the operations research community. We focus on personnel rostering and scheduling problems in which there are often past schedules available and show that it is possible to automatically learn constraints from such data. To realize this, we adapted some techniques from the constraint programming community and we have extended them in order to cope with multi-dimensional data. The method uses a tensor representation of the data, which helps in capturing the dimensionality as well as the structure of the data, and applies tensor operations to find the constraints that are satisfied by the data. To evaluate the proposed algorithm, we used constraints from the Nurse Rostering Competition and generated solutions that satisfy these constraints; these solutions were then used as data to learn constraints. Experiments demonstrate that the proposed algorithm is capable of producing human readable constraints that capture the underlying characteristics of the data.

-----

* 1482
  Value of product location information in agrifood logistics decision making: a multi-level perspective
  Viet Nguyen, Jacqueline Bloemhof

Product location information (PLI) generated by tracking &amp; tracing systems enables a continuous monitoring of material flows and logistics operations across the supply chains. In agrifood sector, tracking &amp; tracing systems are often associated with implications on product safety and quality, whereas the values from logistics management perspective have been overlooked. These values can be seen from a multi-level decision making perspective and can be further increased by big data analytics methods. Using the tracking &amp; tracing systems from a crossdocking warehouse in the Dutch floriculture sector as an illustrative case, we extensively discuss the value of historical and real-time PLI in supporting logistics decisions at different levels, i.e. strategic, tactical and operational. Currently, workforce scheduling is the most challenging decision at the crossdocking due to the time constraints and the wide daily and hourly fluctuations in inbound product volumes. Employing data-mining based big-data approaches, we develop a framework to integrate the use of historical and real-time PLI in making strategic, tactical and real-time workforce plans. This paper helps managers gain a systematic understanding of the multi-level values and opportunities enabled by PLI and tracking &amp; tracing systems. Moreover, it stresses on the necessity to accelerate the implementation of big data analytics across the supply chain functions to make further uses of PLI and big data from other systems.

-----

* 2102
  Data, information, knowledge and optimisation.
  Patrick De Causmaecker

Implicit information in data, explicit knowledge expressed apart from the actual optimisation models all could contribute to the quality of the decision support. In present practice, however, the use of this information and knowledge is very limited. Use of data often targets specific purposes and explicit knowledge on the context can certainly not always be taken into account by present optimisation techniques. Some illustrative examples demonstrate that this may be a missed opportunity. We review existing methods, techniques and procedures presently allowing to incorporate data and knowledge to improve the decision support in accuracy and stability. On-line as well as off-line decision can profit from existing tools. We identify opportunities for improvement as well as some open questions.

Session 154: TD-09: Evaluation as a Service for Optimization and Data Science
Chair(s): Szymon Wasik, Maciej Antczak

-----

* 1004
  Optil.io: platform supporting evaluation as a service architecture
  Szymon Wasik, Maciej Antczak, Jan Badura, Artur Laskowski

Reliable and trustworthy evaluation of algorithms is a challenging process. Firstly, each algorithm has its strengths and weaknesses, and the selection of test instances can significantly influence the assessment process. Secondly, the measured performance of the algorithm highly depends on the test environment architecture, i.e., CPU model, available memory, cache configuration, operating system&#39;s kernel, and even compilation flags. Finally, it is often difficult to compare algorithm with software prepared by other researchers, because their algorithms are not available as open source, do not compile or do not work correctly for all test cases.

Evaluation as a Service (EaaS) is a cloud computing architecture that tries to make assessment process more reliable by providing online tools and test instances dedicated to the evaluation of algorithms.

One of such platforms is Optil.io which gives the possibility to define problems, store evaluation data and evaluate solutions submitted by researchers in almost real time. Researchers can upload their algorithms in the form of binaries or the source code that is compiled online. Then the solution is executed and evaluated at the server, in the homogenous runtime environment and the live ranking of all submitted algorithms is presented. Optil.io is dedicated, but not limited, to optimization algorithms and allows for organization of programming challenges.

-----

* 2947
  Using contests to crack algorithmic and data science problems
  Rinat Sergeev, Jin Paik

The mission of our Lab is spurring the development of a science of innovation through a systematic program of solving real-world technical challenges while simultaneously conducting rigorous scientific research and analysis. On the development side, we design and field competitions that generate the best computer code and data analytics for NASA, other Government Agencies, and Academia. The real-world problems are formulated as open-door challenges, where developers are awarded prizes for the production of finished packages that can be delivered at comparatively low cost. On the academic side, we are studying the process of crowdsourcing from the problems that can be most efficiently solved by the crowd, to incentives for crowd members to participate in the solution development, to contest mechanics that optimizes the effort of the crowd.
In my talk, I am going to discuss:
- The strong sides and trade-offs of crowdsourcing;
- The niche of data science and algorithmic competitions;
- The methodology of problem selection and adaptation for competitions;
- The role, importance, and trade-offs of automatic evaluation and progress feedback in competitions;
- Practical examples of automatic evaluation systems driving extreme value outcomes of competitions in diverse areas - from space to healthcare.

-----

* 3104
  Review of online platforms supporting evaluation of algorithms
  Artur Laskowski, Maciej Antczak, Jan Badura, Szymon Wasik

There are plenty of online platforms that support evaluation of algorithms, such as Codeforeces, TopCoder, CheckIO, Optil.io, and many others. Such platforms have to provide a reliable, homogeneous runtime environment which compares algorithm implementations in precisely same conditions in a cloud. They can support a limited number of programming languages or a plethora of programming technologies. They can be open sourced or commercial. They can provide thousands of problems that users can solve or just give the possibility to upload self-designed content.

During this presentation, we would like to classify online platforms supporting evaluation of

algorithms based on their applications, such as the organization of competitive programming challenges, facilitating teaching, providing tools for solving data mining challenges, or compiling and executing code in the cloud. During our work that we want to present, we have reviewed over one hundred of such platforms and compared them regarding their scientific and industrial potential. Furthermore, we proposed the formal definition of online judge system and summarized commonly used evaluation methodologies. During this short presentation, we would like to inspire everybody who is dealing with evaluation of algorithms to consider using online judge platforms.

-----

* 2903
  Brilliant challenges: lessons learned while collecting evaluation data for optimization problems
  Maciej Antczak, Jan Badura, Artur Laskowski, Jacek Blazewicz, Szymon Wasik

Brilliant Challenges contest aimed to collect interesting, applicable, optimization problems that could have been published at the Optil.io platform and addressed by its users. Participants submitted problems related to classic areas of combinatorial optimization (i.e., knapsack problem, Steiner tree) but also to its modern applications such as Internet shopping optimization and even problems that require computationally complex simulation to calculate the objective function’s value (i.e., simulation of cellular automaton). During this presentation, we will shortly describe these problems and the winners of the competition. Moreover, we will present the main shortcomings observed in evaluation data submitted by our participants demonstrating how difficult it is to prepare data and scripts allowing for reliable evaluation of algorithms.

Session 155: WA-09: Graphs, Data and Optimization
Chair(s): Pieter Leyman

-----

* 2095
  A scale-independent structural definition of community detection in graphs
  Alexander Braekevelt, Pieter Leyman, Patrick De Causmaecker

Finding groups of connected individuals or communities in graphs has received considerable attention in literature. In an attempt to quantify the intuitive concept of “strongly connected individuals with few outgoing connections”, many definitions have been proposed. Most of these, however, provide little insight in the quality and properties of the resulting subgraphs in general graphs. Applications in OR might require more rigorous definitions. We propose to analyze different characteristics from literature, derived from clique relaxations, and determine a better definition of connectedness. An important aspect in this regard is the scalability of the definition, which we explicitly take into account. We subsequently discuss our algorithm, based on the Moody &amp; White approach for finding account. We subsequently discuss our algorithm, based on the Moody and White approach for finding vertex connected sets, to detect the defined types of communities, given a minimum and maximum value. Hence, the proposed technique is able to identify groups of nodes which satisfy predefined properties. In terms of data, we use both existing (real-world) data and fictitious data generated ourselves. Finally, we also analyze the impact of data parameters on algorithm performance, and consider the link with instance difficulty. Possible applications include social media network analysis (e.g. Facebook, Twitter), bioinformatics and real-world vehicle routing.

-----

* 2127
  Network structure and dynamics of European stock markets
  José G. Dias

The analysis of co-movement and contagion between stock markets has been based mostly on the computation of correlation-based measures on raw data. A disadvantage of using this type of measures is that they are sensitive to extreme observations and do not take time dependency into account. This paper introduces a new method that filters time dependence and data outliers by using hidden Markov models prior to the computation of association measures. Then, networks analysis is performed to understand cross-relations between stock markets. The application to European stock markets illustrates the improved understanding of dynamics using different measures of association and dynamic visualization techniques, in particular rolling measures of association, hierarchical trees, and minimum spanning trees (MST).

-----

* 2706
  Supporting evaluation of algorithms during pace challenges
  Jan Badura, Maciej Antczak, Artur Laskowski, Szymon Wasik

The goal of the Parameterized Algorithms and Computational Experiments (PACE) Challenge is to investigate the applicability of algorithmic ideas studied and developed in the subfields of multivariate, fine-grained, parameterized, or fixed-parameter tractable algorithms. The 2017 and 2018 editions of PACE were hosted using Optil.io platform. Up to this point, the organizers run solutions submitted by users only once, at the end of the contest, using manually executed scripts. Optil.io platform provided them with automatic, continuous evaluation method. Contestants could submit their algorithms to the platform and have them assessed almost instantly using 200 instances. 100 of them were public ones, for which participants could see their results in real time and compare those results with others. Another 100 instances were private. Results of evaluation using these instances were

visible only to organizers, and after the challenge deadline, they were used to determine the winner. For each instance algorithm could use up to 30 minutes of processor time, thus requiring to perform 100 hours of computation per submission. Thanks to the parallelization, contestants could check their results on public instances after just around 90 minutes. Uploading problems from both editions to Optil.io platform was possible with a minimal additional effort from the organizers, and after minimal pitfalls, at the beginning of 2017 edition, the evaluation worked seamlessly.

-----

* 2273
  A self-adaptive evolutionary algorithm for the intermittent traveling salesman problem with different temperature profiles
  Pieter Leyman, Patrick De Causmaecker

We study the intermittent traveling salesman problem (ITSP), which extends the traditional TSP by including temperature restrictions for each node. These restrictions limit the allowable consecutive visiting time for each node, which leads to each node possibly requiring multiple visits. We extend previous work by considering different types of temperature profiles, which determine the temperature increase and decrease rates. We propose an evolutionary algorithm (EA) and show the contribution of each of the EA's individual components. Especially the analysis of different solution representations, which is often overlooked in literature, is of great importance. The effect of including the EA's parameters in the solution representation, allowing for self-adaptive parameter control is studied as well, and compared with adaptive rules which adjust parameter values based on solution quality. Additionally, the consequence of introducing a learning component in the EA's mutation operator, is investigated. Based on the test results, we draw conclusions regarding the contribution of each of the proposed algorithm components, and show the impact of the different temperature profiles.

Session 156: WB-09: Data Science in Optimization Algorithms
Chair(s): Andrew J. Parkes, Daniel Karapetyan

-----

* 3557
  Quick termination of poor performers in automated algorithm configuration
  Daniel Karapetyan, Andrew J. Parkes, Thomas Stützle

Much of automated algorithm configuration is concerned with choosing the best algorithm out of a given set. This process is expensive as it usually requires full-scale testing of each algorithm, i.e. running the algorithm for a specified amount of time on several instances. Naturally, a poor performer can be detected earlier, saving computational resources. A well-developed approach (racing) is concerned with terminating after testing the algorithm on a subset of the test instances. Instead, in this talk, we study another approach focusing on termination of each of the test runs early. Our method is based on collecting the data along the run of the tests. We train our system to terminate poor performers as soon as the signal is strong enough to make the decision. We demonstrate a simple strategy that significantly reduces the overall experiment time while still finding 99% of the top algorithms. In future, we hope to use more advanced data science techniques to exploit the rich data that can be collected, to further improve the efficiency of the method. In our experiments, we use the Conditional Markov Chain Search framework, as it is specifically designed for automated generation of combinatorial optimisation algorithms.

-----

* 3124
  Features of genetic algorithms realization for the Euclidean combinatorial optimization problems
  Oleksii Kartashov

Let C be an arbitrary combinatorial set. We carry out a one-to-one mapping of C onto some subset E of the arithmetic Euclidean space. The set obtained as a result of such a mapping is called Euclidean combinatorial set. The specific properties of Euclidean combinatorial sets allow us to propose original approaches to the realization of genetic algorithms (GA) for solving optimization problems on these sets. As an individual (gene), we will consider the sequence of coordinates of the point of E. The crossover operator is of the greatest interest in this GA implementation. One of the most common ways of a crossover is to exchange of a parts from parent individuals. Unfortunately the resulting descendants may not belong E. A natural approach to obtain valid descendants is the solution of the problem of projecting an arbitrary point onto the set E. The solution of the problem for various classes of Euclidean combinatorial sets is known. Developing this approach it is possible to propose the other ways of obtaining descendants. For example, a new descendant may be obtained as a linear combination of parent individuals with coefficients determined by the values of the goal function at these points. The better the value, the greater the ratio. The proposed approaches were realized for the problems of optimal placement of a set of circles of different radii without mutual intersections, which was reduced to the optimization problem on set of permutations.

-----

* 3962
  Metaheuristics for the search for SAT heuristics
  Andrew Burnett, Andrew J. Parkes

For local search to be effective it often requires internal heuristics in order to decide which moves

such be made. Designing such heuristics is often the task of an expert, and can require significant efforts and time. Hence, it is a natural goal to automate the creation of such heuristics. This has previously been addressed by genetic programming (and other evolutionary methods). In this talk, we will discuss the potential for metaheuristics to create heuristics. The specific context is that of propositional satisfiability (SAT) and a set of local search methods known as `WalkSAT'. We will discuss the potential for machine learning methods to enhance this search process.

-----

* 3969
  Machine learning of the quality of machine-generated heuristics
  Andrew J. Parkes, Ender Özcan, Asghar Neema Mohammad Beglou

Previous work has shown that it is possible to automatically generate heuristics for the online bin packing problem. Furthermore, these heuristics significantly outperform human generated heuristics. However, there are two outstanding challenges: the search for heuristics can be slow as it is a form of simulation or expensive optimisation; the heuristics have no evident explanation, and this limits resulting scientific insights and generalisations. In this work, we study potential usages of machine learning to provide quickly computed surrogates and also to provide explanations of the space of heuristics.

Session 250: WC-09: Integrating Machine Learning in Optimization Methods
Chair(s): Kevin Tierney

-----

* 950
  A data science approach for OD matrix analysis in transportation modelling
  Lídia Montero, Jaume Barceló, Xavier Ros-Roca

Origin-Destination (OD) matrices are a main input to most of the transport analysis procedures. Travel demand analysis is the first step in the traditional 4 steps Transport Models, and it is usually based on a well-established methodology, traditionally consisting on detailed travel household surveys, accurately designed, supported by careful sampling procedures, whose results are combined with socioeconomic census data. However, the resulting matrices still need to go through additional checking and validation processes before being used for transport analysis, especially when they will become a major input to transport models. A variety of checking and validation process have been proposed, nevertheless there still remain many alternatives to explore. The first objective of the research reported is to propose a methodological approach to systematically analyze the spatio-temporal structure of the trip patterns represented by modal OD matrices based on the use of Principal Components Analysis and Correspondence Analysis. The main methodological contribution in the validation process consists on considering analysis of similarities in the global mobility pattern inter-districts and/or inter-neighborhoods of common knowledge among mobility analysts. To do this a Correspondence Analysis (CA) is applied to the daily modal mobility matrix in the city of Barcelona. Hourly pattern analysis is addressed using Principal Component Analysis (PCA).

-----

* 3860
  Optimization of Shows Schedules on Linear Television
  Sebastian Souyris, Jaime Miranda

The hyper-competitive live and video entertainment industry offers to consumers an array of high-quality products to choose from like never before. At the same time, content providers must make decisions carefully in order to grow or maintain a profitable position. At a tactical level, decision makers must decide what shows to create and acquire in order to maximize the medium-term ratings. At an operational level, schedulers must program the shows in order to maximize the short-term target demographic viewership. In this talk, we present rating forecasting and show scheduling models that in conjunction lift network audience. We use gradient boosting trees to predict audience and an integer programming model that use patterns to schedule shows. Our models are used by leading US television networks. We present the approach and results.

-----

* 2279
  Online algorithm selection for operational-level combinatorial optimisation problems
  Hans Degroote, Lars Kotthoff, Jose Luis Gonzalez-Velarde, Bernd Bischl, Patrick De Causmaecker

The classic approach to solving an NP-HARD problem is to develop a specialised algorithm. We apply a more high-level approach: to collect a portfolio of algorithms that solve the problem well and to learn when to use which algorithm while solving problem instances. This is called online algorithm selection. The core idea is to use a supervised learning method to create a regression model for each algorithm, predicting its performance. To do so, a set of descriptive features is required. For each new instance, the predicted best algorithm is selected. Its performance is then observed and used to further refine its regression model, resulting in better selections over time. An advantage of online algorithm selection is that no time must be spent on developing and fine-tuning new algorithms: it automatically learns to combine the algorithms already developed. Furthermore, new algorithms can be added to the portfolio, and the method can learn when to use them. Online algorithm selection is most useful for problems at the operational level, as those must be solved often (generating a lot of new performance data) and in a short time (if much time is available, all algorithms can simply be run in parallel). Simulations on the ASlib algorithm selection benchmark show that the proposed method consistently manages to learn over time. It performs especially well when starting from some initial training data (warm start), but can also outperform the single best algorithm without.

-----

* 1533
  Deep learning assisted heuristic tree search for the container pre-marshalling problem
  André Hottung, Shunji Tanaka, Kevin Tierney

One of the key challenges for operations researchers solving real-world problems is designing and implementing high-quality heuristics to guide their search procedures. In the past, machine learning techniques have failed to play a major role in operations research approaches, especially in terms of guiding branching and pruning decisions. We integrate deep neural networks into a heuristic tree search procedure to decide which branch to choose next and to estimate a bound for pruning the search tree of an optimization problem. We call our approach Deep Learning assisted heuristic Tree Search (DLTS) and apply it to a well-known problem from the container terminals literature, the container pre-marshalling problem (CPMP). Our approach is able to learn heuristics customized to the CPMP solely through analyzing the solutions to CPMP instances, and applies this knowledge within a heuristic tree search to produce the highest quality heuristic solutions to the CPMP to date.

Session 451: WD-09: Optimization of Machine Learning Models
Chair(s): Dimitri Papadimitriou

-----

* 2381
  Fractional Euclidean distance matrices for Kriging surrogate model
  Natalija Pozniak, Leonidas Sakalauskas

The paper deals with the application of fractional distance matrices to construct the Kriging surrogate model. The objective is to discuss and demonstrate the implementation of the Kriging surrogate modelling technique and verify the computational efficiency of the method. Some examples are tested by computer simulation to illustrate the effectiveness of the proposed model when compared to the Kriging and Shepard extrapolator. The comparison results verify the computational efficiency. Once constructed, the Kriging surrogate model can be used for data extrapolation and optimization. The surrogate model developed has been applied for modeling of surface wastewater treatment filters. The effectiveness of filters filled with construction waste and biocarbon was analysed.

-----

* 2383
  Recurrent parameter estimation algorithm in hidden Markov models with application to multivariate data analysis and signal recognition
  Jūratė Vaičiulytė, Leonidas Sakalauskas

Hidden Markov Models (HMMs) have been extensively used in dynamic systems modeling as well as applied to numerous practical areas. Thus, we present a recurrent algorithm for online hidden Markov model parameter estimation and its application to multivariate data clustering and classification. In this work, we focus on continuous HMMs and assume that model parameters have the multivariate normal distribution with unknown mean vector and covariance matrix. We aim to find optimal model parameter estimates. Therefore, the maximum likelihood method was used to estimate the unknown parameters of the model, and formulas for the adapted recurrent Expectation-Maximization (EM) algorithm of HMMs parameter estimation were derived. The complexity of the adapted recurrent EM algorithm is linear in respect of a sample size. Hence, the parameters of the HMM model are updated with each new observation received, without remembering the previous observation set. The adapted recurrent EM algorithm consists of two main parts – model training and recognition. Therefore, we can continuously estimate model parameters and perform online recognition at the same time. In our experiment, the implemented algorithm was used for multivariate data clustering and adapted for isolated word recognition. The results of the experiment are discussed in this work as well.

-----

* 3027
A novel optimization based algorithm for multi-class data classification problem
Fatih Rahim, Metin Turkay

Multi-class data classification is a supervised machine learning problem that involves assigning data to multiple groups. We present a novel MILP-based algorithm that splits each class’s data set into subsets such that the subsets of different classes are linearly separable. At each iteration, we form a subset of samples out of the set of unassigned samples by a MILP model that maximizes the cardinality of the new subset. The generated subset which is linearly separable from the subsets of other classes is removed from the unassigned sample set. The rest of the samples are used for generating new subsets and the algorithm terminates when all the samples are assigned. There is a hyperplane that separates each pair of subsets of different classes such that the hyperplanes form a polyhedral region for each subset and the regions of different classes are disjoint. We build classifiers based on the convex hulls of the subsets and the polyhedral regions for the testing phase. We evaluate our approach on benchmark problems and compare it with the methods from the literature. We conclude that our optimization-based algorithm complemented with the proposed classifiers, provides competitive results in terms of prediction accuracy.

-----

* 2086
Combined structure and weight learning for neural networks
Dimitri Papadimitriou

Compelling arguments have emerged over time for using multi-layer neural networks (NN) as general pattern for solving supervised machine learning problems. However, designing and training the right NN for a given learning task remains subject to many theoretical gaps and practical concerns. Most algorithms focus on finding the weight values that minimize the empirical loss given a NN structure provided as input to the minimization problem, that is nonlinear, nonconvex and high-dimensional. Common techniques to cope with efficiency (number of layers/depth, units/size, etc.) still rely on handling the NN structure as hyperparameter that is tuned using a validation set or on the dropout method which samples from exponential number of different thinned structures to reduce overfitting at the detriment of increasing training time. Instead, combining structure and weight learning yields a multi-objective minimization problem, where the NN structure itself becomes a variable of the problem (instead of a parameter). The proposed method for solving the combined problem extends the Generalized Benders (variable decomposition) technique to non-convex objective functions. It proceeds by i) iteratively building/tuning the structure of the NN over which the weight learning algorithm performs and ii) adapting this structure according to the learning rate gain and to the epsilon-accuracy of the solution produced.