Tabular Survey: Surrogate Models in Combinatorial Optimization

Version 5

Martin Zaefferer, Thomas Bartz-Beielstein

May 9, 2017

1 Introduction

Surrogate models are typically used to lighten the burden of costly objective function evaluations in real-world optimization. Here, we focus on surrogate models in the discrete, combinatorial optimization domain. This encompasses, e.g., the following data representations: mixed integer, integer (ordinal or categorical), binary, permutation, string, tree and graph. After a short explanation, this document provides a tabular overview of the literature.

While the goal is to give a broad overview, the table is unlikely to be complete. Please be so kind and email the author[1] if you see any shortcomings. Additions, corrections or comments are welcomed and encouraged. Currently, approaches with probabilistic / distribution-based models (e.g., Estimation of Distribution algorithms) are not included, or only included as optimizers, not as models. Since they may be included in the future, feel free to notify the author of any interesting candidates.

2 Explanations

Table 1 presents a tabulated overview of the literature on discrete, surrogate-model based optimization. Table 2 collects works that are also of interest in this context, but which only deal with the modeling aspect and not with the optimization aspect. Specified are the data types, modeling strategies, model types, optimizers, and some further information about the solved problems. The following six strategies of dealing with discrete data structures are listed:

1. The naive approach: As long as the data can still be represented as a vector (binary variables, integers, categorical data, permutations) the modeling technique may simply ignore the discrete structure, and work as usual.
2. Custom modeling: A specific modeling solution is tailored to fit the needs of a certain application.
3. Inherently discrete models: Use of models that are inherently discrete. One example is the use of tree-based models, like regression trees, random forests or, in some cases, artificial neural networks.
4. Mapping: Discrete variables or structures may be mapped to a more easily handleable representation. Examples for this approach are the random key mapping for permutations or dummy variables for categorical variables.
5. Feature extraction: Instead of directly modeling the relation between an object (or its representation) and its quality, it is possible to calculate real-valued features of the objects. E.g., some properties of a tree or graph can be extracted (path lengths, tree depths, etc.). These numeric features can than be modeled with standard techniques.
6. Similarity-based modeling: Where available, measures of (dis)similarity may be used to replace continuous measures that are, e.g., employed in similarity-based models like k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Radial Basis Function Networks (RBFN) or Kriging.

These strategies are not necessarily mutually exclusive. Some methods may either combine several strategies, or else, can be classified as belonging to several strategies.

[1] martin.zaefferer@gmx.de

Table 1: Column *data* lists data type: ordinal integer (ord), categorical integer (cat), binary (bin), permutation (per), signed permutation (-per), trees (tre), other (oth). Column *strategy* lists the modeling strategy, as introduced in sec. 2. NA indicates that a strategy does not clearly fit into any category. Column cost refers to the (time, material, etc.) cost per objective function evaluation. Budget is the maximum number of allowed evaluations (or in some cases a time limit). If no budget is specified, another stopping criterion was used instead. Where applicable, column *dimension* lists the dimensionality of the problem, i.e., the number of variables. Abbreviations: Generalized Linear Model (GLM), Non-dominated Sorting Genetic Algorithm II (NSGA2), Covariance Matrix Adaption Evolution Strategy (CMA-ES), Evolution Strategy (ES), Genetic Algorithm (GA), Differential Evolution (DE), Simulated Annealing (SA), Artificial Neural Networks (ANN), Ant Colony Optimization (ACO), Radial Basis Function Networks (RBFN), Support Vector Machine (SVM), Branch and Bound (B&B), Multi-start Local Search (MLS), k-Nearest Neighbor (k-NN). Question marks indicate that the respective information was not found in the given reference.

data	strategy	model	optimizer	cost	budget	dimension	topics	reference
mix, cat, ord	1, 3	Kriging, Tree	visual, statistical analysis	high	≤ 100	2, 9	parameter tuning	Bartz-Beielstein and Markon [2004]
mix, bin	NA	Kriging	B&B	low	?	6	benchmark	Davis and Ierapetritou [2007]
mix, ord	1,3	RBFN	grid search	low	few hundred, thousand	2-11	benchmark	Holmström [2007]
mix, ord, cat	6	RBFN	ES	low / ∼high	560 / 280	15 / 23	benchmark, medical image analysis	Li et al. [2008]
mix, ord, bin	1, 6	Kriging	B&B	high	50 - 500	5-18	electrical engineering, water management	Hemker [2008]
mix, ord, cat	3, 4	Linear Regression, Tree	sampling	low / ∼high	-	2-13	algorithm tuning	Bartz-Beielstein [2009]
mix, ord, cat	3, 6	Random Forest, Kriging	MLS	∼high	-	4-76	algorithm tuning	Hutter et al. [2010]
mix, bin, cat	6	RBFN+cluster	GA	low	2,000	12	benchmark, real-world: chemical industry	Bajer and Holeňa [2010]
mix, ord, cat	4+6	RBFN+cluster+ GLM	GA	low	several thousand	4-13	benchmark, real-world: chemical industry	Bajer and Holeňa [2013]
mix, ord, cat	4	SVM	NSGA2	?	2,000	10	finite element, multi criteria	Herrera et al. [2014]
mix, ord	1	RBFN, Kriging	sampling, GA	low	500	2-60	benchmark	Müller [2015]
mix, cat	1, 3	Random Forest, Kriging, Kernel Regression	MLS, EDA / CMA-ES	high	200-500	14-36	benchmark	Eggensperger et al. [2015]
mix, ord	1	Kriging	DE	low / high	1000-2000	4-20	benchmark, chip design	Liu et al. [2016]
mix, ord, cat	3	Random Forest	focus search	low – high	300	8	machine learning, tuning, multi criteria	Horn and Bischl [2016]
ord	1/3	ANN, SVR, RBFN, Kriging	GA	low	500	56	benchmark, noise	Horng and Lin [2013]
ord	1	RBFN	EDA	low	10 minutes	6	heuristic selection	Martins et al. [2017]
bin	3	Finite State Machine	GA	low	$100,000$	$1,000$	performance testing	Corne et al. [2003]
bin	1/3	ANN	SA	high	?	16	real world, pump positioning	Rao and Manju [2007]
bin	6	RBFN	GA	low	$\dim^2$	10-25	NK-Landscape	Moraglio and Kattan [2011a]

bin	6	RBFN	GA	high	100	10-40	benchmark, package deal negotiation	Fatima and Kattan [2011]
bin	6	Kriging	GA	low	dim^2	10-25	NK-Landscape	Zaefferer et al. [2014b]
bin	3	Bayesian ANN	sampling	low	thousands	512	Chemical Data	Hernández-Lobato et al. [2016]
-per	2	custom	brute force	high	28	6	signed permutation, real world: weld sequence	Voutchkov et al. [2005]
per	6	RBFN	GA	low	100	30 - 32	benchmark	Moraglio et al. [2011]
per	6	Kriging	GA	low	100	12 - 32	benchmark	Zaefferer et al. [2014b]
per	6	Kriging	GA	low	200	10 - 50	distance selection	Zaefferer et al. [2014a]
per	6	Kriging	ACO	low	100 - 1,000	50 - 100	benchmark, tuning	Pérez Cáceres et al. [2015]
per	6	RBFN	GA	instance dependent	1,000	50 - 1,928	numerical stability, real world: cell suppression	Smith et al. [2016]
per	6	Kriging	brute force, GA	low	100	5 - 10	kernel definiteness	Zaefferer and Bartz-Beielstein [2016]
tre	6	RBFN	GA	low	100		symbolic regression	Moraglio and Kattan [2011b]
tre	5,6	k-NN	GA	high	30,000		phenotypic similarity, genetic programming	Hildebrandt and Branke [2015]
tre	5	k-NN	GA	high	55,000		job shop scheduling, genetic programming	Nguyen et al. [2014]
tre	NA	RBFN	GA	low	100		symbolic regression, parity	Kattan and Ong [2015]
tre	2,5	k-NN	GA	high	25,000		job shop scheduling, genetic programming	Nguyen et al. [2016]
tre	5	Random Forest	GA	low	15,000		benchmark, genetic programming	Pilát and Neruda [2016]
oth	6	k-NN	GA	rather low	20,000 - 200,000	161 - 259	real-world: protein structure	Custódio et al. [2010]
oth, bin	3	Landscape State Machine	GA	low	several thousand	10, 100	protein sequence, NK-Landscape	Rowe et al. [2010]
oth	6	Kriging	GA	high	few hundreds		graph-based, real-world, protein structure	Romero et al. [2013]
oth	6	Kriging	grid search	high	40	23	mixed hierarchical variables, ANN tuning	Swersky et al. [2013]
oth	5	ANN	DE	low	several hundreds	40 - 500	assignment problem, dynamic	Hao et al. [2016]
oth	6	Kriging	gradient based	high	1000	120	automatic chemical design	Gómez-Bombarelli et al. [2016]
oth, mix, cat	3	RF	focus search	low - high	200	2 - 35	parameter tuning	Bischl et al. [2017]

Table 2: Column *data* lists data type: Mixed variables (mix), ordinal integer (ord), categorical integer (cat), binary (bin), permutation (per), signed permutation (-per), trees (tre), other (oth). Column *strategy* lists the modeling strategy, as introduced in sec. 2. NA indicates that a strategy does not clearly fit into any category. Where applicable, column *dimension* lists the dimensionality of the problem, i.e., the number of variables.

data	strategy	model	dimension	topics	reference
mix, cat	6	Kriging	1,8	kernels, data center temperature model	Qian et al. [2008]
mix, cat	6	Kriging	2,8	kernels, computational fluid dynamics	Zhou et al. [2011]
mix, cat, oth	6	Kriging		mixed variables, hierarchical variables: graph structure	Hutter and Osborne [2013]
mix, cat	4,6	Kriging		mixed variables	Duvenaud [2014]

References

Lukáš Bajer and Martin Holeňa. Surrogate model for continuous and discrete genetic optimization based on rbf networks. In *Intelligent Data Engineering and Automated Learning – IDEAL 2010*, volume 6283 LNCS, pages 251–258, 2010.

Lukáš Bajer and Martin Holeňa. Surrogate model for mixed-variables evolutionary optimization based on glm and rbf networks. In *SOFSEM 2013: Theory and Practice of Computer Science*, volume 7741 LNCS, pages 481–490, 2013.

Thomas Bartz-Beielstein. Sequential parameter optimization. In Jürgen Branke, Barry L. Nelson, Warren Buckler Powell, and Thomas J. Santner, editors, *Sampling-based Optimization in the Presence of Uncertainty*, number 09181 in Dagstuhl Seminar Proceedings, Dagstuhl, Germany, 2009. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany. URL `http://drops.dagstuhl.de/opus/volltexte/2009/2115`.

Thomas Bartz-Beielstein and Sandor Markon. Tuning search algorithms for real-world applications: a regression tree based approach. In *Proceedings of the 2004 Congress on Evolutionary Computation*, Portland, OR, US, 2004. Institute of Electrical and Electronics Engineers (IEEE). doi: 10.1109/cec.2004.1330986. URL `http://dx.doi.org/10.1109/CEC.2004.1330986`.

Bernd Bischl, Jakob Richter, Jakob Bossek, Daniel Horn, Janek Thomas, and Michel Lang. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *arXiv*, 2017. URL `http://arxiv.org/abs/1703.03373`.

David Corne, Martin Oates, and Douglas Kell. Landscape state machines: Tools for evolutionary algorithm performance analyses and landscape/algorithm mapping. In *Proceedings of the 2003 International Conference on Applications of Evolutionary Computing*, EvoWorkshops'03, pages 187–198, Berlin, Heidelberg, 2003. Springer-Verlag. ISBN 3-540-00976-0. URL `http://dl.acm.org/citation.cfm?id=1765642.1765662`.

Fábio L Custódio, Helio JC Barbosa, and Laurent Emmanuel Dardenne. Full-atom ab initio protein structure prediction with a genetic algorithm using a similarity-based surrogate model. In *Proceedings of the Congress on Evolutionary Computation (CEC'10)*, pages 1–8, New York, NY, USA, 2010. IEEE.

Eddie Davis and Marianthi Ierapetritou. A kriging based method for the solution of mixed-integer nonlinear programs containing black-box functions. *Journal of Global Optimization*, 43(2-3):191–205, aug 2007. doi: 10.1007/s10898-007-9217-2. URL `http://dx.doi.org/10.1007/s10898-007-9217-2`.

David K. Duvenaud. *Automatic Model Construction with Gaussian Processes*. PhD thesis, University of Cambridge, 2014.

Katharina Eggensperger, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Efficient benchmarking of hyperparameter optimizers via surrogates. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 1114–1120. AAAI Press, 2015. ISBN 0-262-51129-0. URL `http://dl.acm.org/citation.cfm?id=2887007.2887162`.

Shaheen Fatima and Ahmed Kattan. Evolving optimal agendas for package deal negotiation. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, GECCO '11, pages 505–512, New York, NY, USA, 2011. ACM.

Rafael Gómez-Bombarelli, David K. Duvenaud, José Miguel Hernández-Lobato, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *CoRR*, abs/1610.02415, 2016.

Jing-hua Hao, Min Liu, Jian-hua Lin, and Cheng Wu. A hybrid differential evolution approach based on surrogate modelling for scheduling bottleneck stages. *Computers & Operations Research*, 66:215–224, 2016.

Thomas Hemker. *Derivative Free Surrogate Optimization for Mixed-Integer Nonlinear Black Box Problems in Engineering*. PhD thesis, Technische Universität Darmstadt, December 2008.

J. M. Hernández-Lobato, E. Pyzer-Knapp, A. Aspuru-Guzik, and R. P. Adams. Distributed thompson sampling for large-scale accelerated exploration of chemical space. In *NIPS Workshop on Bayesian Optimization*, Barcelona, Spain, 2016.

Manuel Herrera, Aurore Guglielmetti, Manyu Xiao, and Rajan Filomeno Coelho. Metamodel-assisted optimization based on multiple kernel regression for mixed variables. *Structural and Multidisciplinary Optimization*, 49(6):979–991, 2014.

Torsten Hildebrandt and Jürgen Branke. On using surrogates with genetic programming. *Evolutionary Computation*, 23(3):343–367, Jun 2015.

Kenneth Holmström. An adaptive radial basis algorithm (ARBF) for expensive black-box global optimization. *Journal of Global Optimization*, 41(3):447–464, nov 2007. doi: 10.1007/s10898-007-9256-8. URL `http://dx.doi.org/10.1007/s10898-007-9256-8`.

Daniel Horn and Bernd Bischl. Multi-objective parameter configuration of machine learning algorithms using model-based optimization. Technical Report Nr. 68/2016, TU Dortmund, 2016.

S.-C. Horng and S.-Y. Lin. Evolutionary algorithm assisted by surrogate model in the framework of ordinal optimization and optimal computing budget allocation. *Information Sciences*, 233:214–229, 2013.

F. Hutter, H. H. Hoos, and K. Leyton-Brown. Sequential model-based optimization for general algorithm configuration (extended version). Technical Report TR-2010-10, University of British Columbia, Department of Computer Science, 2010. Available online: http://www.cs.ubc.ca/~hutter/papers/10-TR-SMAC.pdf.

Frank Hutter and Michael A. Osborne. A kernel for hierarchical parameter spaces. Technical Report eprint arXiv:1310.5738, ARXIV, 2013.

Ahmed Kattan and Yew-Soon Ong. Surrogate genetic programming: A semantic aware evolutionary search. *Information Sciences*, 296:345–359, 2015.

Rui Li, M. T M Emmerich, J. Eggermont, E. G P Bovenkamp, T. Bäck, J. Dijkstra, and J. Reiber. Metamodel-assisted mixed integer evolution strategies and their application to intravascular ultrasound image analysis. In *Proceedings of the Congress on Evolutionary Computation (CEC'08)*, pages 2764–2771, New York, NY, USA, 2008. IEEE.

Bo Liu, Nan Sun, Qingfu Zhang, Vic Grout, and Georges Gielen. A surrogate model assisted evolutionary algorithm for computationally expensive design optimization problems with discrete variables. In *2016 IEEE Congress on Evolutionary Computation (CEC)*. Institute of Electrical and Electronics Engineers (IEEE), jul 2016. doi: 10.1109/cec.2016.7743986. URL `http://dx.doi.org/10.1109/CEC.2016.7743986`.

Marcella S. R. Martins, Mohamed El Yafrani, Myriam R. B. S. Delgado, Markus Wagner, Belaïd Ahiod, and Ricardo Lüders. Hseda: A heuristic selection approach based on estimation of distribution algorithm for the travelling thief problem. In *Proceedings of the Genetic and Evolutionary Computation Conference 2017*, page 8, Berlin, Germany, july 2017. ACM.

Alberto Moraglio and Ahmed Kattan. Geometric generalisation of surrogate model based optimisation to combinatorial spaces. In *Proceedings of the 11th European Conference on Evolutionary Computation in Combinatorial Optimization*, EvoCOP'11, pages 142–154, Berlin, Heidelberg, Germany, 2011a. Springer.

Alberto Moraglio and Ahmed Kattan. Geometric surrogate model based optimisation for genetic programming: Initial experiments. Technical report, University of Birmingham, 2011b.

Alberto Moraglio, Yong-Hyuk Kim, and Yourim Yoon. Geometric surrogate-based optimisation for permutation-based problems. In *Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*, GECCO '11, pages 133–134, New York, NY, USA, 2011. ACM.

Juliane Müller. MISO: mixed-integer surrogate optimization framework. *Optimization and Engineering*, 17(1):177–203, jun 2015. doi: 10.1007/s11081-015-9281-2. URL `http://dx.doi.org/10.1007/s11081-015-9281-2`.

Su Nguyen, Mengjie Zhang, Mark Johnston, and Kay Chen Tan. Selection schemes in surrogate-assisted genetic programming for job shop scheduling. In *Simulated Evolution and Learning, 10th International Conference, SEAL*, pages 656–667. Springer Science + Business Media, 2014. doi: 10.1007/978-3-319-13563-2_55. URL http://dx.doi.org/10.1007/978-3-319-13563-2_55.

Su Nguyen, Mengjie Zhang, and Kay Chen Tan. Surrogate-assisted genetic programming with simplified models for automated design of dispatching rules. *IEEE Transactions on Cybernetics*, pages 1–15, 2016. doi: 10.1109/tcyb.2016.2562674.

Leslie Pérez Cáceres, Manuel López-Ibáñez, and Thomas Stützle. Ant colony optimization on a limited budget of evaluations. *Swarm Intelligence*, pages 1–22, 2015.

Martin Pilát and Roman Neruda. Feature extraction for surrogate models in genetic programming. In *Parallel Problem Solving from Nature – PPSN XIV*, pages 335–344. Springer Nature, 2016. doi: 10.1007/978-3-319-45823-6_31. URL http://dx.doi.org/10.1007/978-3-319-45823-6_31.

Peter Z. G Qian, Huaiqing Wu, and C. F. Jeff Wu. Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics*, 50(3):383–396, 2008.

S. V. N. Rao and S. Manju. Optimal pumping locations of skimming wells. *Hydrological Sciences Journal*, 52(2):352–361, 2007.

Philip A Romero, Andreas Krause, and Frances H Arnold. Navigating the protein fitness landscape with Gaussian processes. *Proceedings of the National Academy of Sciences*, 110(3):E193–E201, 2013.

William Rowe, David C. Wedge, Mark Platt, Douglas B. Kell, and Joshua Knowles. Predictive models for population performance on real biological fitness landscapes. *Bioinformatics*, 26(17):2145–2152, jul 2010. doi: 10.1093/bioinformatics/btq353. URL http://dx.doi.org/10.1093/bioinformatics/btq353.

Jim Smith, Christopher Stone, and Martin Serpell. Exploiting diverse distance metrics for surrogate-based optimisation of ordering problems: A case study. In *Proceedings of the 2016 on Genetic and Evolutionary Computation Conference*, GECCO '16, pages 701–708, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4206-3. doi: 10.1145/2908812.2908854.

Kevin Swersky, David Duvenaud, Jasper Snoek, Frank Hutter, and Michael Osborne. Raiders of the lost architecture: Kernels for bayesian optimization in conditional parameter spaces. In *NIPS workshop on Bayesian Optimization in Theory and Practice*, 2013.

I. Voutchkov, A.J. Keane, A. Bhaskar, and Tor M. Olsen. Weld sequence optimization: The use of surrogate models for solving sequential combinatorial problems. *Computer Methods in Applied Mechanics and Engineering*, 194(30-33):3535–3551, Aug 2005.

Martin Zaefferer and Thomas Bartz-Beielstein. Efficient global optimization with indefinite kernels. In *Parallel Problem Solving from Nature–PPSN XIV*, pages 69–79. Springer, 2016.

Martin Zaefferer, Jörg Stork, and Thomas Bartz-Beielstein. Distance measures for permutations in combinatorial efficient global optimization. In Thomas Bartz-Beielstein, Jürgen Branke, Bogdan Filipič, and Jim Smith, editors, *Parallel Problem Solving from Nature–PPSN XIII*, pages 373–383, Cham, Switzerland, 2014a. Springer.

Martin Zaefferer, Jörg Stork, Martina Friese, Andreas Fischbach, Boris Naujoks, and Thomas Bartz-Beielstein. Efficient global optimization for combinatorial problems. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation*, GECCO '14, pages 871–878, New York, NY, USA, 2014b. ACM.

Qiang Zhou, Peter Z. G. Qian, and Shiyu Zhou. A simple approach to emulation for computer models with qualitative and quantitative factors. *Technometrics*, 53(3):266–273, 2011.