Data Driven Evolutionary Optimization of Complex Systems: Big Data Versus Small Data

Yaochu Jin
Department of Computer Science
University of Surrey
Guildford, GU2 7XH, United Kingdom
yaochu.jin@surrey.ac.uk

ABSTRACT

This invited talk Existing evolutionary algorithms typically assume that there are explicit objective functions available for fitness evaluations. In the real world, such explicit objective functions may not be available in many cases. For example, many industrial optimization problems such as structural design [8] need to perform computationally very intensive numerical simulations, such as computational fluid dynamic simulations or finite element analysis, where a large set of partial differential equations must be solved. In many process industry optimization problems, no explicit models exist for describing the relationship between the final quality of the product and the decision variables, such as temperature and humidity. Thus, only historical experimental data can be used for optimization. There are also cases where only factual data can be collected. A good example of such optimization problems is trauma systems design [11], where only patient records are available for optimization.

For solving such optimization problems, evolutionary optimization can be conducted only using a data-driven approach. The main challenges in data-driven evolutionary optimization can roughly be divided into two categories according to the amount of available data, namely, small data and big data. The lack of data can mainly be attributed to the fact that data acquisition is very expensive, either computationally or costly. In [11], data-driven evolutionary optimization problems are divided into two paradigms, one termed off-line data-driven optimization, where no new data can be actively sampled, and the other on-line data-driven optimization, where a small number of new data points can be collected.

For small data driven evolutionary optimization, the use of surrogate techniques to assist evolutionary algorithms becomes indispensable [6, 7]. In particular in on-line data-driven surrogate-assisted evolutionary optimization, two important questions arise. The first question is, which surrogate model should be used, and second, when a new data should be collected and where. Many model management techniques have been developed for answering these questions.

Surrogate techniques can be categorized into deterministic models, such as polynomials, artificial neural networks, radial-basis-function neural networks, and probabilistic models such as Gaussian processes, also known as Kriging models. The biggest benefit of probabilistic models is that these models can provide a confidence level in addition to a predicted objective value. Such information about uncertainty is particularly helpful in model management in determining when and where a new data should be sampled [5].

Surrogates can be categorized from a different perspective, e.g., whether a surrogate aims to approximate the local fitness landscape or the global landscape, in relation to the distribution of the current population. It has been shown that a proper combination of global and local models can effectively accelerate the search process [9, 10]. The key issue here is how to choose samples properly to train the local and global surrogates, and how to combine the local and global surrogate models.

In contrast to small data driven evolutionary optimization, there are cases where large amount of data is involved for optimization, which may also be subject to noise and uncertainty [12]. In such cases, the key question is how to manipulate the data properly so that useful and sufficient information can be extracted while minimising the amount of data for calculating the objective functions to reduce the computational time. One intuitive idea is to cluster the data into groups so that representative data can be use. An essential question is how to adaptively tune the number of clusters so that the computational cost can be reduced without negatively influencing the search process. One such idea was reported in [11] where a regression function is built to learn the relationship between the cluster number and the maximum error that will not mislead the non-dominated sorting based selection in multi-objective optimization. Many questions remain open in this line of research, in particular when different selection criteria are used in evolutionary algorithms.

In spite of the great access achieved in data-driven and surrogate-assisted evolutionary optimization, a few important challenges remain to be addressed. First, little efforts have been reported on surrogate-assisted optimization of high-dimensional and large scale optimization problems. In the literature, the highest dimension in surrogate-assisted evolutionary optimization is 50 and most surrogate-assisted evolutionary algorithms can solve problems having around ten decision variables. This is partly due to curse of di-

GECCO'16 Companion, July 20–24, 2016, Denver, CO, USA.

ACM ISBN 978-1-4503-4323-7/16/07.
DOI: http://dx.doi.org/10.1145/2908961.2931715

mensionality, which means that a large number of data is needed for training surrogates having reasonably good accuracy, and partly due to the extremely high computational cost for constructing surrogates such as Gaussian processes. To solve these problems, ideas proposed in big data research could be used [3, 4].

Second, data-driven optimization of combinatorial or mix-integer optimization problems deserve more research attention. The main challenge in data-driven optimization of combinatorial or mix-integer optimization problems results from the fact that surrogates are indispensable in both small data and big data driven optimization and no efficient surrogate modelling techniques exist for combinatorial and mix-integer problems [2].

Third, more attention should be paid to data-driven surrogate-assisted optimization of constrained problems. Preliminary work indicates that these problems are more difficult to solve in the sense that surrogate training may become trickier when some of the training data are infeasible solutions [1].

Finally, it is extremely challenging to solve off-line data-driven optimization problems where only small amount of data is available and no new data can be actively sampled during the optimization process. Since there is no easy way to validate the optimal solutions obtained by the surrogate-assisted search, it is sometimes intractable to verify the effectiveness of search algorithm before it is employed for solving the real problem.

1. ACKNOWLEDGEMENTS

This work was supported in part by an EPSRC grant (No. EP/M017869/1) and the Joint Research Fund for Overseas Chinese, Hong Kong and Macao Scholars of the National Natural Science Foundation of China (No. 61428302).

2. REFERENCES

[1] T. Chugh, K. Sindhya, K. Miettinen, J. Hakanen, and Y. Jin. On constraint handling in surrogate-assisted evolutionary many-objective optimization. In *Problem Solving from Nature*, 2016 (Submitted).

[2] E. Davis and M. Ierapetritou. A kriging based method for the solution of mixed-integer nonlinear programs containing black-box functions. *Journal of Global Optimization*, 43:191âĂŞ205, 2009.

[3] J. Hensman and N. Lawrence. Gaussian processes for big data through stochastic variational inference. In *NIPS Workshop on Big Learning*, 2012.

[4] T. N. Hoang, Q. M. Hoang, and K. H. Low. A unifying framework of anytime sparse gaussian process regression models with stochastic variational inference for big data. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015.

[5] D. Horn, T. Wagner, D. Biermann, C. Weihs, and B. Bischl. Model-based multi-objective optimization: Taxonomy, multi-point proposal, toolbox and benchmark. In *Evolutionary Multi-Criterion Optimization*, LNCS 9018, pages 64–78, 2015.

[6] Y. Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing*, 9(1):3–12, 2005.

[7] Y. Jin. Surrogate-assisted evolutionary computation: recent advances and future challenges. *Swarm and Evolutionary Computation*, 1(2):61–70, 2011.

[8] Y. Jin and B. Sendhoff. A systems approach to evolutionary multiobjective structural optimization and beyond. *IEEE Computational Intelligence Magazine*, 4(3):62–76, 2009.

[9] D. Lim, Y. Jin, Y.-S. Ong, and B. Sendhoff. Generalizing surrogate-assisted evolutionary computation. *IEEE Transactions on Evolutionary Computation*, 14(3):329–355, 2010.

[10] C. Sun, Y. Jin, J. Zeng, and Y. Yu. A two-layer surrogate-assisted particle swarm optimization algorithm. *Soft Computing*, 19(6):1461–1475, 2015.

[11] H. Wang, Y. Jin, and J. O. Jansen. Data-driven surrogate-assisted multi-objective evolutionary optimization of a trauma system. *IEEE Transactions on Evolutionary Computation*, 2016.

[12] Z.-H. Zhou, N. Chawla, Y. Jin, and G. Williams. Big data opportunities and challenges: Discussions from data analytics perspectives. *IEEE Computational Intelligence Magazine*, 9(4):62–74, 2014.