Summary of 'Automatic composition and optimization of multicomponent predictive systems' written by Manuel Martin Salvador, Marcin Budka and Bogdan Gabrys.

The machine learning results mainly depends on the quality and quantity of used data and selected learning model. The typical workflow consists of following steps: data gather, data pre-processing (including missing value handling, outlier detection and handling, data transformation and dimensionality reduction), data sampling, choosing of a model, training, cross validation and final deployment of trained model.

In typical cases the data pre-processing can be handled by the scientific experts depending on the domain, however the labour-instensiveness of this task is too high to be even considered as feasible by human. Although the supervision of the experts in many cases is welcome. Considering the number of pre-processing tasks, including each own optimization, inclines scientists to develop more automatic solution of this very labor-intensive work. There are quite a lot attempts being made to optimize particular problems, including building of programming frameworks, although those underlying approaches are lacking the context of cross-domain generalization.

The next step in the machine learning workflow is to find the optimal algorithm which performs significantly better for a given pre-processed dataset. This task also can be very time and computationally intensive as there are multiple alternative models to be checked. This step is also followed by two actions, which assure the quality of a model - multiple hyperparameter optimization and cross validation of the results. The training and validation of each model for a given set of hyperparameters makes the obtaining of the best solution almost impossible without using metaheuristics and significant simplifications. The entire problem of finding the optimal model with optimized hyperparameters is very challenging.

The Authors of 'Automatic composition and optimization of multicomponent predictive systems' have noticed that generally in the literature a comparison of different models bases on the data always preprocessed in the same way. They have concluded that some of the models are performing better if the data is pre-processed with them in mind. The optimization of the hyperparameters can also be included as a part of finding the model as the process of choosing the model and optimization is very similar - different parametrization of the same model can be treated as a different solution and in fact model. The model selection and hyperparameters optimization is a joint problem and is known in a literature as CASH (Combined Algorithm Selection and Hyperparameter optimization) problem.

In the approach taken, in the mentioned earlier article, authors include the pre-processing steps as a part of the whole CASH problem.This approach is called by the authors Multicomponent Predictive System (MCPS). By combining both processes together the complexity of the problems grow, but also provides more significantly reduction in the amount of time and also increases the accuracy of the solution, beating human experts. The search space of the CASH problem can be seen as a grid of possible solutions. The size of such grid is too high to use exhaustive search algorithms and compute all the solutions. One of the proposed and used in some of the industries approaches is a ROAR (Random Online Aggressive Search), which simplifies the and narrows the large search spaces. The authors for their experiments decided to limit the length of the MCPS and explicitly set the boundaries on the execution time and memory usage. The mathematical foundation and the way of formalizing the MCPS base on the Petri nets (PN). The process of building MCPS is typically iterative, labor and knowledge intensive, hence the authors say that reliable fully automated approach still doesn't exist.

The extensive experiments done by the authors proves that it is possible to automate the composition of multicomponent predictive systems. Based on the mathematical foundation of MCPS, authors claim that a formal verification of correctly composed model is still non-trivial problem. The obtained results indicate that random search performs worse than sequential model-based optimization for the same given time of optimization. The exceptional cases were noticed for the datasets coming from the process industry. The fine-tuning of the hyperparameters of solution might not be the best strategy for finding the best solution in the search space and it is suggested to exploit the search space by increasing the limits of the simulation so that more models can be validated.

Mateusz Gasior