AutoML: an appealing challenge to stimulate research in data mining automation

Marc Boullé

Orange Labs - 22300 Lannion - France

I am a senior researcher in the data mining research group of Orange Labs. My main research interests include statistical data analysis, especially data preparation and modeling for large databases. In Orange, there are many requests for data mining studies, in a wide diversity of application domains and tasks, structure and scale of data, constraints, resource or business requirements. The bottleneck to a wide spread of data mining solutions is the lack of data analysts. One solution to leverage this problem is the automation of the data mining process, with a focus on the following criterions:

- genericity, to address a wide range of cases,
- no parameter, for automation purpose,
- robustness, to produce reliable predictions,
- accuracy, to obtain good predictions,
- understandability, to get interpretable models (e.g. in the marketing field),
- computational efficiency, with good scalability in train and/or deployment,
- resource management, with efficient use of available resources (IO, RAM, CPU...).

Following these objectives, I developed regularized methods for feature preprocessing, feature selection, feature construction for multi-table data mining, correlation analysis via k-coclustering, model averaging of selective naive Bayes classifiers and regressors. Mainly, these methods exploit large model families (for feature construction, preprocessing, selection...) with parsimonious Bayesian prior defined on all models and efficient optimization heuristic to retrieve the most probable models given the data. This Bayesian model selection approach provides both automation and statistical reliability, with guarantee that information in feature is optimally detected.

The AutoML Challenge is very relevant to stimulate research towards automation of data mining. This is a key feature to address the long tail of projects with small to medium business impact and to help data analysts in the few complex projects with large business impact. This challenge is well in line with my research interests. It is an appealing way to evaluate my methods, at least on some of the criterions summarized above. I then designed some python scripts to follow the challenge settings, and could exploit Khiops, the Orange Labs automatic tool for mining large databases. The same tool was used throwout the challenge and obtained consistently competitive results. In future work, to go beyond the limits of the selective naive Bayes classifier, I will consider more powerful classifiers to improve accuracy while decreasing as little as possible the performance on the other evaluation criterions, such as computational efficiency or interpretability. Beyond this, the main interest of the challenge is to set a milestone in the path to data mining automation and to assess the next steps. This raises numerous interesting open problems:

- given a task, data, objectives per criterion, constraints, computational resources, how to automatically and efficiently build a solution?
- beyond the case of accuracy, how to assess the performance on other criterions such as understandability?
- in case of multiple criterions, how to exploit the Pareto front of all the optimal solutions?
- how to efficiently extend automation to data beyond the tabular format?

I expect that the AutoML challenge will stimulate the research and practice, and invite the data mining community to propose novel solutions.