1 The Orange Labs Software for Mining Large Databases

We summarize the participation of Orange Labs to the AutoML challenge, both to evaluate the Orange Labs data mining tool and to provide a competitive reference for all challenge participants.

1.1 Presentation

In Orange, the main french telecommunication operator, 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...).

There does not exist a single solution that simultaneously optimizes each objective. The issue is then to find a Pareto optimal solution with good trade-offs between the competing objectives. Given this multi-objective optimization problem, Orange Labs has developed a tool name Khiops[1], that is fully automatic, scalable, highly robust, and obtains competitive performance on the other criterions. In an industrial context like the Orange telecommunication operator, 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.

1.2 Machine Learning Algorithms

Khiops exploits regularized methods for variable preprocessing, variable selection, variable construction for multi-table data mining, correlation analysis via k-coclustering, model averaging of selective naive Bayes classifiers and regressors. We summarize below the Selective Naive Bayes (SNB) classifier introduced in (Boullé, 2007). It extends the Naive Bayes classifier owing to an optimal estimation of the class conditional probabilities, a Bayesian variable selection and a Compression-based Model Averaging.

Optimal preprocessing. Numerical variables are preprocessed using supervised discretization to evaluate the class conditional probabilities. In the MODL approach (Boullé, 2006), the discretization is turned into a model selection problem and solved in a Bayesian way. Using a hierarchical prior distribution on the discretization parameters, the Bayes formula is applicable to derive an exact analytical criterion to evaluate the posterior probability of a discretization model. A 0-1 normalized version of this criterion provides a univariate informativeness evaluation of each input variable. Similarly, categorical variables are preprocessed using supervised value grouping (Boullé, 2005). This Bayesian model selection approach provides both automation and statistical reliability, with guarantee that information in input variables is optimally detected.

Bayesian Approach for Variable Selection. The naive independence assumption can harm the performance when violated. In (Boullé, 2007), the Selective Naive Bayes (SNB) classifier (Langley and Sage, 1994) is trained using a Bayesian model selection approach to select the best subset of variables (Guyon et al., 2006). Efficient search heuristics with super-linear computation time are proposed, on the basis of greedy forward addition and backward elimination of variables.

[1] Available at `www.khiops.com`

Compression-Based Model Averaging. Instead of taking the best subset of variables, the method introduced in (Boullé, 2007) averages all the classifiers resulting from different subsets of variables, using a logarithmic smoothing of the posterior distribution of the trained classifiers. The weighting scheme on the models reduces to a weighting scheme on the variables, and finally results in a single Naive Bayes classifier with weights per variable.

Extension to regression. The same framework is extended to regression in (Hue and Boullé, 2007). Input variables are preprocessed by bivariate discretization models, that optimize jointly the discretization (or value grouping) of the input and target variables, so as to estimate the univariate conditional probability of the rank of the target values. Then, the SNB regressor combines the univariate estimators to get a multivariate predictor. It is noteworthy that the SNB rank regressor is able to predict the whole conditional probability distribution of the target ranks given the input values. For point estimation as in usual regression settings, the expectation of the target value is computed by integrating over the estimated target rank probability distribution.

1.3 Adaptation to AutoML Settings

The Khiops tool was used throughout the challenge, using python scripts to be compliant to the challenge settings. Beyond the necessary but easy adaptation to the input/output requirements, the python scripts also had to manage the sparse format, the any-time learning settings and the scoring metrics.

Sparse datasets. Since Khiops doesn't support sparse arrays at the moment, the sparse datasets were recoded using a multi-table format, with a root table containing an instance Id and the class value, and a secondary table in zero to many relation ship, with one record per available value and three variables: instance Id, variable Id and variable value. The sparse variables were also sorted by decreasing frequencies in a preprocessing step, so as to focus on the most frequent values. This way, the sparse datasets could be stored as sparse files, then processed as usual by Khiops, with possibily a focus on the most frequent values.

Anytime learning. The Khiops tool aims at exploiting all the available data so as to produce a solution with good performance on all criterions presented above. This can take a long time for large datasets, and the python scripts were written to launch Khiops in a series of scenarios of increasing sizes. The smallest size is 10,000 instances, chosen randomly, 1000 variables, chosen randomly for the dense datasets and by decreasing frequencies for the sparse datasets, and 100 preprocessed variables as input of the SNB predictor, chosen by decreasing univariate informativeness. A model is trained using the smallest size and iteratively trained again with twice the numbers of instances, variables and preprocessed variables, as long as enough training time is still available.

Scoring metrics. Khiops standard outputs are target probabilities, or target values in case of point estimation. Minimum adaptation was used for the scoring metrics. For the BAC and F1 scores, the SNB output probabilities were computed by replacing the initial target prior probabilities by artificially balanced prior probabilities.

1.4 Analysis of the Challenge Results

The same tool was used throughout the challenge and obtained consistently competitive results. However, whereas Khiops is driven to perform equally well on all the evaluation criterions presented above, the challenge has a strong focus on the accuracy criterion, and the SNB results are often strongly dominated by alternative approaches. Actually, the SNB classifier is resilient to noise and to redundancies between the input variables, but it is blind to non-trivial interactions between the variables. This can be leveraged by feature engineering, relying on domain expertise rather than

on statistical expertise. More accurate classification methods are available, such as random forests, gradient boosting methods, support vector machines or neural networks. However, these methods require intensive feature engineering to get a flat input data table representation, are prone to over-fitting, are mainly black-box, not suitable for an easy interpretation of the models and finally require fine parameter tuning, both time consuming and expertise intensive. Future work is then necessary to incorporate more powerful classifiers in Khiops. The issue is extend automation to problems where excellent accuracy is needed, while decreasing as little as possible the performance on the other evaluation criterions, such as computational efficiency or interpretability.

Beyond this first straightforward objective, the main interest of the AutoML 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?

Hopefully, the AutoML challenge will stimulate the research and practice and incite the data mining community to propose novel solutions.

References

Boullé, M. (2005). A Bayes optimal approach for partitioning the values of categorical attributes. *Journal of Machine Learning Research*, 6:1431–1452.

Boullé, M. (2006). MODL: a Bayes optimal discretization method for continuous attributes. *Machine Learning*, 65(1):131–165.

Boullé, M. (2007). Compression-based averaging of selective naive Bayes classifiers. *Journal of Machine Learning Research*, 8:1659–1685.

Guyon, I., Gunn, S., Nikravesh, M., and Zadeh, L. (2006). *Feature Extraction: Foundations And Applications*. Springer.

Hue, C. and Boullé, M. (2007). A new probabilistic approach in rank regression with optimal bayesian partitioning. *Journal of Machine Learning Research*, pages 2727–2754.

Langley, P. and Sage, S. (1994). Induction of selective Bayesian classifiers. In *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Morgan Kaufmann.