Experimental Evaluation of the e-LICO Meta-Miner

(Extended Abstract)

Phong Nguyen and **Alexandros Kalousis** and **Melanie Hilario** [1]

1 Introduction

Operator selection is the task of selecting the right operator for building not only valid but also optimal data mining (DM) workflows in order to solve a new learning problem. One of the main achievements of the EU-FP7 e-LICO project[2] has been to develop an *Intelligent Data-Mining Assistant* (IDA) to assist the DM user in the construction of such DM workflows following a cooperative AI-planning approach [2] coupled with a new meta-learning approach for mining past DM experiments, referred as the e-LICO meta-miner [3]. The idea of meta-mining [1] is to build meta-mined models from the full knowledge discovery process by analysing learning problems and algorithms in terms of their characteristics and core components within a declarative representation of the DM process, the Data Mining OPtimization ontology (DMOP)[3].

In this paper, we provide experimental results to validate the e-LICO meta-miner's approach to the operator selection task. We experimented on a collection of real-world datasets with feature selection and classification workflows, comparing our tool with a default strategy based on the popularity of DM workflows. The results show the validity of our approach; in particular, that our selection approach allows to rank appropriately DM workflows with respect to the input learning problem. In the next section, we briefly review the meta-miner. In section 3, we present our results. And in section 4, we conclude.

2 The e-LICO Meta-Miner

The role of the AI-planner is to plan valid DM workflows by reasoning on the applicability of DM operators at a given step i according to their pre/post-conditions. However, since several operators can have equivalent conditions, the number of resulting plans can be in the order of several thousands. The goal of the meta-miner is to select at a given step i among a set of candidate operators A_i the k best ones that will optimize the performance measure associated with the user goal g and its input meta-data $\mathbf{m}$ in order to gear the AI-planner toward optimal plans. For this, the meta-miner makes use of a quality function Q which will score a given plan w by the quality q of the operators that form w as:

$$Q(w|g,\mathbf{m}) = q^*(o_1|g,\mathbf{m}) \prod_{i=2}^{|\mathcal{T}(w)|} q(o_i|\mathcal{T}(w_{i-1}),g,\mathbf{m}) \quad (1)$$

where $\mathcal{T}(w_{i-1}) = [o_1,..,o_{i-1}]$ is the sequence of previous operators selected so far, and q^* is an initial operator quality function. Thus the meta-miner will qualify a candidate operator by its conditional probability of being applied given all the preceding operators, and select those that have maximum quality to be applied at a step i. In order to have reliable probabilities, the meta-miner makes use of frequent workflow patterns extracted from past DM processes with the help of the DMOP ontology such that the operator quality function q is approximated as:

$$q(o|\mathcal{T}(w_{i-1}),g,\mathbf{m}) \approx \text{aggr}\left\{\frac{\text{supp}(f_i^o|g,\mathbf{m})}{\text{supp}(f_{i-1}|g,\mathbf{m})}\right\}_{f_i^o \in F_i^o} \quad (2)$$

where aggr is an aggregation function, F_i^o is the set of frequent workflow patterns that match the current candidate workflow w_i^o built with a candidate operator o, and f_{i-1} is the pattern prefix for each pattern $f_i^o \in F_i^o$. More importantly, the quality of a candidate workflow w_i^o will depend on the support function $\text{supp}(f_i^o|g,\mathbf{m})$ of its matching patterns. As described in [3], this support function is defined by learning a dataset similarity measure which will retrieve a dataset's nearest neighbors Exp_N based on the input meta-data $\mathbf{m}$. We refer the reader to [3] for more details. In the next section, we will deliver experimental results to validate our meta-mining approach.

3 Experiments

To meta-mine real experiments, we selected 65 high-dimensional biological datasets representing genomic or proteomic microarray data. We applied on these bio-datasets 28 feature selection plus classification workflows, and 7 classification-only workflows, using tenfold cross-validation. We used the 4 following feature selection algorithms: Information Gain, *IG*, Chi-square, *CHI*, ReliefF, *RF*, and recursive feature elimination with SVM, *SVMRFE*; we fixed the number of selected features to ten. For classification we used the 7 following algorithms: one-nearest-neighbor, *1NN*, the *C4.5* and *CART* decision tree algorithms, a Naive Bayes algorithm with normal probability estimation, *NBN*, a logistic regression algorithm, *LR*, and SVM with the linear, *SVM_l* and the rbf, *SVM_r*, kernels. We used the implementations of these algorithms provided by the RapidMiner data mining suite with their default parameters. We ended up with a total of $65 \times (28 + 7) = 2275$ base-level DM experiments, on which we gathered all experimental metadata; folds predictions and performance results, dataset metadata and workflow patterns, for meta-mining [1].

We constrain the AI-planner so that it generates feature selection and/or classification workflows only. We did so in order for the past experiments to be really relevant for the type of workflows we want to design. Note that the AI-planner can also select from operators with which we have not experimented. These are for feature selection, Gini Index, *Gini*, and Information Gain Ratio, *IGR*. For classification, we used a Naive Bayes algorithm with kernel-based probability

[1] University of Geneva, Switzerland, email: Phong.Nguyen@unige.ch
[2] http://www.e-lico.eu
[3] The DMOP is available at http://www.dmo-foundry.org

estimation, *NBK*, a Linear Discriminant Analysis algorithm, *LDA*, a Rule Induction algorithm, *Ripper*, a Random Tree algorithm, *RDT*, and a Neural Network algorithm, *NNet*.

3.1 Baseline Strategy

In order to assess how well our meta-miner performs, we need to compare it with some baseline. To define this baseline, we will use as the operators quality estimates simply their frequency of use within the community of the RapidMiner users. We will denote this quality estimate for an operator o by $q_{def}(o)$. Additionaly, we will denote the quality of a DM workflow, w, computed using the $q_{def}(o)$ quality estimations by $Q_{def}(w)$, thus:

$$Q_{def}(w) = \prod_{o_i \in \mathcal{T}(wf)} q_{def}(o_i) \qquad (3)$$

The score $q_{def}(o)$ focuses on the individual frequency of use of the DM operators, and does not account for longer term interactions and combinations such as the ones captured by our frequent patterns. It reflects thus simply the popularity of the individual operators. In what concerns the most frequently used classification operators, these were *C4.5*, followed by *NBN*, and *SVM_l*. For the feature selection algorithms, the most frequently used were *CHI* and *SVM-RFE*.

3.2 Evaluation and Comparison Strategy

The evaluation will be done in a leave-one-dataset-out manner, where we will use our selection strategies on the remaining 64 datasets to generate workflows for the dataset that was left out. On the left-out dataset, we will then determine the K best workflows using the baseline strategy as well as using the meta-miner selection strategy. To compare the performance of the ordered set of workflows constructed by each strategy, we will use the average estimated performance of the K workflows on the given dataset, which we will denote by ϕ_a. We will report the average of ϕ_a over all the datasets. Additionally, we will estimate the statistical significance of the number of times over all the datasets that the meta-miner strategy has a higher ϕ_a than the baseline strategy; we will denote this by ϕ_s. We estimated the neighborhood Exp_N of a dataset using $N = 5$ nearest neighbors. We will compare the performance of the baseline and of the meta-miner for $K = 1, 3, 5$ generated workflows in order to have a large picture of their overall performance.

3.3 Performance Results and Comparisons

K=1. The top-1 workflow selected by the baseline strategy is *CHI-C4.5*. When we compare its performance against the performance of the top-1 workflow selected by the meta-miner given in the first row of table 1, we can see that the meta-mining strategy gives an average performance improvement of around 6% over the baseline strategy. In addition, its improvement over the baseline is statistically significant in 53 datasets over 65, while the baseline wins only on 11 datasets.

K=3. The two other workflows selected by the baseline strategy additionally to the top-1 are *CHI-NBN* and *CHI-SVM_l*. When we extend the selection to the three best workflows, we obtain the results given in the second row of table 1, where we see that the average predictive performance improvement over the baseline strategy is around 2%. As before, the meta-miner achieves significantly better performance than the baseline in a larger number of baselines datasets than vice-versa.

K=5. The two other workflows selected by the baseline strategy additionally to the top-3 are *SVMRFE-C4.5* and *SVMRFE-SVM_l*. We give the results of the five best workflows selected by the meta-miner in the last row of table 1, where we observe similar trends as before; 2% of average performance improvement and statistical difference in the number of improvement in favor of the meta-mining strategy.

		ϕ_a	ϕ_s	
$K = 1$	Q_{def}	71.92%	11/65	
	Q	77.68%	53/65	p=2e-7
$K = 3$	Q_{def}	75.04%	22/65	
	Q	77.28%	41/65	p=0.046
$K = 5$	Q_{def}	75.18%	18/65	
	Q	77.14%	44/65	p=0.006

Table 1. Performance results and comparisons for the top-K workflows.

3.4 Selected Workflows

We will briefly discuss the top-K workflows selected by the meta-miner. For $K = 1$, we have on a plurality of datasets the selection of the *LDA* classifier, an algorithm we have not experimented with. This happens because within the DMOP ontology this algorithm is related both with the linear, *SVM_l*, and with the NaiveBayes algorithm, both of which perform well on our dataset collection. For $K = 3$ and $K = 5$, we have additionally the selection of the previously unseen *NNet* and *Ripper* classifiers. These operator selections demonstrate the capability of the meta-miner to select new operators based on their algorithm similarities given by the DMOP with past ones.

4 Conclusion and Future Works

This is a preliminary study, but already we see that we are able to deliver better workflow suggestions, in terms of predictive performance, compared to the baseline strategy, while at the same time being able to suggest workflows consisting of operators with which we have never experimented. Future works include more detailed experimentation and evaluation, and the construction of similarity measures combining both the dataset characteristics and the workflow patterns.

ACKNOWLEDGEMENTS

We would like to thank Jörg-Uwe Kietz and Simon Fischer for their contribution in the development and evaluation of the e-LICO meta-miner.

REFERENCES

[1] Melanie Hilario, Phong Nguyen, Huyen Do, Adam Woznica, and Alexandros Kalousis, 'Ontology-based meta-mining of knowledge discovery workflows', in *Meta-Learning in Computational Intelligence*, eds., N. Jankowski, W. Duch, and K. Grabczewski, Springer, (2011).

[2] Jörg-Uwe Kietz, Floarea Serban, Abraham Bernstein, and Simon Fischer, 'Towards Cooperative Planning of Data Mining Workflows', in *Proc of the ECML/PKDD09 Workshop on Third Generation Data Mining: Towards Service-oriented Knowledge Discovery (SoKD-09)*, (2009).

[3] Phong Nguyen, Alexandros Kalousis, and Melanie Hilario, 'A meta-mining infrastructure to support kd workflow optimization', in *Proc. of the PlanSoKD-2011 Workshop at ECML/PKDD-2011*, (2011).