Hybrid Multi-Agent System for Metalearning in Data Mining

Klára Pešková and **Jakub Šmíd** and **Martin Pilát** and **Ondřej Kazík** [1] and **Roman Neruda** [2]

Abstract. In this paper, a multi-agent system for metalearning in the data mining domain is presented. The system provides a user with intelligent features, such as recommendation of suitable data mining techniques for a new dataset, parameter tuning of such techniques, and building up a metaknowledge base. The architecture of the system, together with different user scenarios, and the way they are handled by the system, are described.

1 Introduction

Lately, *data mining* — an automated process of gaining information from datasets — has become an issue of interest in the artificial intelligence. This interest have been whetted by the progress in the computational technology, such as high performance machine clusters or large storage devices, but most importantly by the possibility of an access to enormous amount of data that are collected on daily basis. The datasets vary in many factors as they origin in different areas of human or nature activities. It is hard even for a data mining expert to choose from the wide range of machine learning methods that are used in data mining and to set its parameters to the values that would produce a reasonable results for the specific dataset. Tools that ease up the parameter set up can significantly boost up the productivity of data mining process. Moreover, the automation of the whole process would help those researchers, who are not data mining experts, to enjoy the benefits of this research line. This is where the *metalearning* [3] comes into play.

Metalearning over data mining methods and datasets is a very demanding task, especially with respect to computational performance as it uses results of data mining methods applied on various datasets as its training/testing data. The software that is capable of both data mining and metalearning is by definition a large and complex system. To design the architecture of our system, we have chosen the agent-based approach as it brings many advantages to this complex task. The main one being its distributed and parallel nature — the system can spread over computer networks and be accessed by many users who only by using the system and running their experiments provide the data needed for metalearning algorithms. It also supplies a fast parallel execution of performance demanding tasks. The interconnection of different parts of the system (i.e. the *communication among agents*) is done only by sending messages which results in an easy extensibility and re-usability of the parts of the system — *agents*. It enables researchers to easily add their own components (e.g. custom data mining methods) and to re-use the implemented components in different situations.

We have designed and implemented a multi-agent system (*MAS*) which is capable of executing simple data mining tasks as well as complex metalearning problems (involving not only recommending of data mining methods but also setting their parameters), and it provides all the mechanisms necessary for experimenting with different metalearning approaches. The system is hybrid — it employs combination of different artificial intelligence methods [4].

We use JADE [2] — the multi-agent framework, as a base for our agents; most of the computational agents in our system use Weka [6] data mining methods. The extensibility of our system is assured by the use of the structured ontology language and following the FIPA [1] international standards of agents' communication.

2 Scenarios

To propose an appropriate architecture of our computational MAS, we have considered the following basic scenarios for processing a dataset. In the most simple case the user knows which method and what parameters of this method she would like to use. In the other two basic scenarios, the system uses its intelligent meta-learning features: If the user knows what method to use but does not know how to set its parameters, the system is able to search the parameter space of the method and find a setting that provides good results. In the third case, the user does not even know what method to use and lets the system decide by itself. In this case the system recommends the best possible method or provides a ranking of the methods based on predicted errors and duration. These simple scenarios can be extended into more complex ones — e.g. it is also possible to combine the recommendation of the best method the with parameter space search, when the recommender chosen by the user recommends an interval of the parameter's values.

As a positive side effect, the *metaknowledge base* for metalearning purposes is being build up by each experiment.

3 Role-based Architecture

In order to effectively design our system, we have chosen the organization-centered formalism *AGR* (*Agent-Group-Role*). The *role* is a set of capabilities and responsibilities that the agent accepts by handling the role. *Group* — the building block of a MAS — is a set of agents with allowed roles and interactions, defined by a *group structure*. The multi-agent system then consists of multiple groups which can overlap by agents belonging to more than one group. In this formalism, we abstract from the algorithmic details and inner logic of the agents in the MAS. In our previous work [7], we have

[1] Charles University in Prague, Faculty of Mathematics and Physics, emails: klara@pisecko.cz, jakub.smid@ktiml.mff.cuni.cz, martin.pilat@gmail.com, kazik.ondrej@gmail.com

[2] Institute of Computer Science, Academy of Sciences of the Czech Republic, email: roman@cs.cas.cz

used the ontological formalism of OWL-DL to describe the organizational model.

The following group structures were defined according to the aforementioned scenarios: *administrative group structure*, *computational group structure*, *search group structure*, *recommendation group structure*, *data group structure* and *data-management group structure*.

Our MAS is composed of groups that are instances of these group structures. The architecture is depicted in the Figure 1.

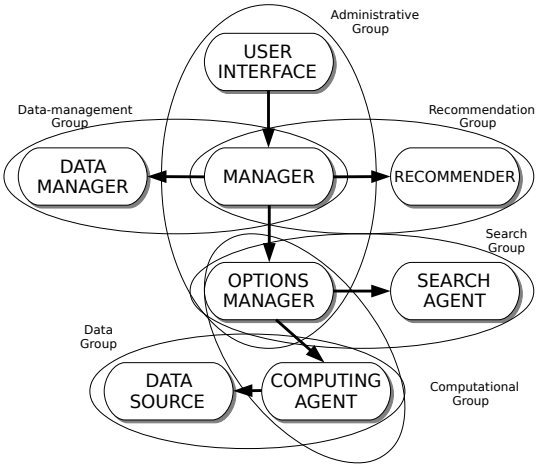

Figure 1. Architecture of our MAS: Group structures

4 Metalearning

The key parts of our system are those providing intelligent metalearning behavior, i.e. agents that provide parameter space search methods and recommender agents. These agents are intended to (at least partially) replace a human expert. They make use of the previous experience gathered by the system, which is captured in the metaknowledge base. It contains results of machine learning experiments and *metadata* — general features of the datasets.

The MAS-based solution allows a flexibility in choice of the parameter space search algorithms, each of these is encapsulated in a search agent. General tabulation, random search, simulated annealing [9], or parallel methods, such as evolutionary algorithms [5], are implemented in our system. Another great benefit of the agent-based approach is the natural capability of parallel execution of computations with various parameters which significantly decreases the time needed for the execution of the parameter space search process.

One of essential features of our MAS is its capability of recommending a suitable computational method for a new dataset, according to datasets similarity and previously gathered experience. The choice of the similar dataset(s) is based on various previously proposed metrics [8], which measure the similarity of their metadata. Our database contains over two million records, that are used to suggest the proper method (including its parameters) and estimate its performance on a new dataset.

The latest version of our MAS contains the following types of recommenders, which differ in the metric used and in the number of recommended methods they provide:

- *Basic recommender* chooses a method based on the single closest dataset using the unweighted metadata metric.
- *Clustering Based Classification* [8] chooses the whole cluster of similar datasets and the corresponding methods, using different sets of metadata features.
- *Evolutionary Optimized Recommenders* are similar to the two above described recommender types, using different weighted metrics, optimized by an evolutionary algorithm.
- *Interval Recommender* recommends intervals of suitable parameter values and leaves their fine-tuning to the parameter space search methods.

Another functionality of our system is a multi-objective optimization of data mining configurations. The search algorithm is employed in order to find beneficial combinations of pre-processings and machine learning methods to the presented data. The minimization is performed in error-rate as well as run-time criteria.

5 Conclusions

In this paper, we presented a multi-agent system for metalearning in data mining, which includes solving of the most important and challenging metalearning tasks – the recommendation of a suitable method for a new dataset, and the tuning of parameters of such methods. We have proposed the systems architecture and proved its usability by an implementation that is used by our research team on a regular basis to conduct metalearning and data mining experiments. The role-based multi-agent approach brings in many advantages into a complex task of metalearning, the main benefit being its easy extensibility. The multi-agent parallel nature of the system speeds up the time consuming tasks significantly.

6 Acknowledgements

Jakub Šmíd and Klára Pešková have been supported by the Charles University Grant Agency project no. 610214, R. Neruda has been supported by the Ministry of Education of the Czech Republic project COST LD 13002.

REFERENCES

[1] The Foundation for Intelligent Physical Agents (FIPA). http://www.fipa.org/.
[2] Java Agent DEvelopment framework. http://jade.tilab.com/.
[3] Pavel Brazdil, Christophe G. Giraud-Carrier, Carlos Soares, and Ricardo Vilalta, *Metalearning - Applications to Data Mining.*, Cognitive Technologies, Springer, 2009.
[4] Oscar Castillo, Patricia Melin, Janusz Kacprzyk, and Witold Pedrycz, *Hybrid Intelligent Systems*, Springer, 2007.
[5] Agoston E. Eiben and J. E. Smith, *Introduction to Evolutionary Computing*, SpringerVerlag, 2003.
[6] M. Hall et al., 'The weka data mining software: An update.', *SIGKDD Explorations*, **11**, (2009).
[7] Ondřej Kazík and Roman Neruda, 'Ontological modeling of meta learning multi-agent systems in OWL-DL', *IAENG International Journal of Computer Science*, **39**(4), 357–362, (Dec 2012).
[8] Ondrej Kazík, Klára Pešková, Jakub Šmíd, and Roman Neruda, 'Clustering based classification in data mining method recommendation', in *ICMLA '13: Proceedings of the 2013 12th International Conference on Machine Learning and Applications*, pp. 356–361, Washington, DC, USA, (2013). IEEE Computer Society.
[9] S. Kirkpatrick, C. D. Gelatt, Jr., and M. P. Vecchi, 'Optimization by simulated annealing', *Science*, **220**, 671–680, (1983).