Machine Learning, 54, 187–193, 2004

Introduction to the Special Issue on Meta-Learning

CHRISTOPHE GIRAUD-CARRIER
ELCA Informatique SA, Ave de la Harpe 22-24, Case Postale 519, CH-1001 Lausanne, Switzerland

RICARDO VILALTA
Department of Computer Science, University of Houston, 4800 Calhoun Rd., Houston TX 77204-3010, USA

PAVEL BRAZDIL
LIACC / Faculty of Economics, University of Porto, R. Campo Alegre, 823, 4150-180 Porto, Portugal

Abstract. Recent advances in meta-learning are providing the foundations to construct meta-learning assistants and task-adaptive learners. The goal of this special issue is to foster an interest in meta-learning by compiling representative work in the field. The contributions to this special issue provide strong insights into the construction of future meta-learning tools. In this introduction we present a common frame of reference to address work in meta-learning through the concept of meta-knowledge. We show how meta-learning can be simply defined as the process of exploiting knowledge about learning that enables us to understand and improve the performance of learning algorithms.

Keywords: meta-learning, meta-knowledge, inductive bias, dynamic bias selection

1. Introduction

The application of Machine Learning (ML) and Data Mining (DM) tools to classification and regression tasks has expanded outside the boundaries of research into the realm of applied research, industry, commerce, and government. Two key aspects play an important role in the successful application of these tools. One is the selection of a suitable predictive model (or combination of models) where expertise is seldom available a priori; users of commercial ML and DM tools must either resort to trial-and-error or expert advice. Clearly, neither solution is completely satisfactory for the end user who wishes to access the technology more directly and cost-effectively. The effectiveness of this process can be enhanced by meta-learning. Meta-learning assistants can provide automatic and systematic user guidance on model selection and method combination.

A second important aspect is how to profit from the repetitive use of a predictive model over similar tasks. The successful application of models in real-world scenarios requires a continuous adaptation to new needs. If a model fails to perform efficiently, one would expect the learning mechanism itself to re-learn, taking into account previous experience. Thus, learning can take place not only at the example (i.e., base) level, but also across tasks (Thrun, 1998; Pratt & Thrun, 1997; Caruana, 1997; Vilalta & Drissi, 2002). Meta-learning capabilities are again needed to control the process of exploiting cumulative expertise gained in the past.

The goal of this special issue is to collect representative work in the field. Recent advances in meta-learning are increasingly filling the gaps in the construction of practical meta-learning assistants and task-adaptive learners, as well as in the development of a solid conceptual framework. Our attempt to systematize the underlying notions in meta-learning has helped us obtain a deeper understanding of this area, including the interaction between the mechanism of learning and the concrete contexts in which meta-learning is applicable. By learning or explaining what causes a learning algorithm to be successful or not on a particular task or domain, we go beyond the (engineering) goal of producing more accurate learners to the (scientific) goal of understanding learning behavior.

Despite the promising direction offered by meta-learning and important recent advances, much work remains to be done. We hope this issue will help convince others in the ML and DM community of the need to invest more effort into this relatively new, interesting subfield of research. There are still new concepts waiting to be discovered which we believe will prove extremely useful in answering both theoretical questions, such as those lying at the heart of statistical learning, and practical questions, related to particular algorithm implementations.

The rest of this introduction to the special issue is organized as follows. In the next section we discuss the topic of meta-learning and meta-knowledge in more detail and their role in different types of learning systems. In Section 3 we review briefly the individual contributions and show how they fit in the scheme suggested earlier. Section 4 presents a short summary and conclusions.

2. Exploiting meta-knowledge in different types of systems

Meta-learning is tightly linked to the process of acquiring and exploiting meta-knowledge. Meta-knowledge can take many different forms and can be defined as any type of knowledge that can be derived in the course of employing a given learning system. Advances in the field of meta-learning hinge around one specific question: how can we acquire and exploit knowledge about learning systems (i.e., meta-knowledge) to understand and improve their performance? To answer this question we need to explain what is meant by a *learning system*. For our purposes, a learning system can be either (1) a single learning algorithm; or (2) a set of different learning algorithms, all employed on the same task. In the following we analyze each case in more detail. Our aim will be to try to examine the particular role of meta-knowledge on different types of learning systems.

2.1. Exploiting meta-knowledge within a single learning algorithm

As we have pointed out earlier, it is important that learning algorithms are able to profit from their repetitive use over similar tasks. Ideally, the models should continuously adapt to new needs. This is usually done by re-learning. Meta-knowledge can capture the cumulative expertise gained on different tasks. The process of re-learning either maintains the learning algorithm unchanged (as in incremental learning), or else enables modifications. Meta-knowledge can be used to control these modifications, which can be either rather simple, such as opting for a particular parameter setting or a particular strategy for parameter

optimization; or more complex, such as is the case with architectures that evolve through experience.

As an example, meta-knowledge can play a role in the process of dynamic selection of inductive bias (Baltes & MacDonald, 1992; Rendell, Seshu, & Tcheng, 1987a, 1987b). A learning algorithm can use meta-knowledge to modify the strength and size of the hypothesis space (DesJardins & Gordon, 1995; Gordon, 1990; Utgoff, 1986). On-line detection of concept drift can help identify contextual clues that allow a learning algorithm to be more selective with respect to training instances for prediction (Widmer & Kubat, 1996; Widmer, 1997).

Also, in the field of inductive transfer and learning-to-learn, meta-knowledge appears in the form of patterns across domains. The general understanding of the nature of patterns across domains is that of invariant transformations. For example, image recognition of a target object is simplified if the object is invariant under rotation, translation, scaling, etc. Hence, learning-to-learn studies how to improve learning by detecting, extracting, and exploiting meta-knowledge in the form of invariant transformation across domains (Thrun & Mitchell, 1995).

2.2. Exploiting meta-knowledge with a set of learning algorithms

Given a set of learning algorithms, we can ask the following: (1) which algorithm is best suited for a given task or application domain? or (2) which is the preferred ordering of the given algorithms? or (3) what is the form of the composite classifier to be employed in the new task?

It has been shown that to answer questions (1) and (2) we need to gather meta-knowledge concerning algorithm performance. The idea is to define a set of domain characteristics or meta-features that are relevant for predicting the performance of the given learning algorithms (Aha, 1992; Michie, Spiegelhalter, & Taylor, 1994; Gama & Brazdil, 1995; Brazdil, 1998; Keller, Paterson, & Berrer, 2000; Brazdil, Soares, & Pinto da Costa, 2003). Meta-features may include information concerning error rates of base-learners, so called landmarkers (Bensusan, & Giraud-Carrier 2000; Pfahringer, Bensusan, & Giraud-Carrier 2000) or the structure of induced decision trees (Bensusan, 1998; Peng et al., 2002). A number of research projects have produced tangible results in this area; prominent examples include the ESPRIT Statlog (1991–1994) and METAL (1998–2001) projects.

The aim may alternatively be to select the best learning algorithm not for the whole dataset, but rather for subareas of the problem domain (Brodley, 1995) or for individual examples (Merz, 1995; Todorovski & Dzeroski, 2003).

There is also ample work on composite classifiers. The process of building a meta-learner from base-learners is known as stacked generalization (Wolpert, 1992). Meta-knowledge is made present through the predictions of level-i generalizers, which are then used to produce higher-level generalizers (Chan & Stolfo, 1998). Here we take the predictions of learning algorithms as relevant information in an attempt to improve the original example representation. Alternatively, one may induce referees which capture the area of expertise of each base learner and arbitrate among them by selecting the most reliable base learner for the examples in each subdomain (Ortega, Koppel, & Argamon, 2001).

In conclusion, the relevance of meta-knowledge as a unifying point demands further work in the characterization of its nature and different manifestations. A clearer picture of the role of meta-knowledge can elucidate the path to building practical meta-learning tools. In the next section we describe the contributions comprised by this special issue using the concept of meta-knowledge as a common frame of reference.

3. Contributions to the special issue

The first contribution to this issue (Soares, Brazdil, & Kuba, 2004) is a meta-learning approach to parameter setting (Section 2.1). The work focuses on Support Vector Machines and the goal is to set the width σ of the Gaussian kernel. The methodology exploits information about past performance of different settings. All past cases (i.e. datasets) are characterized by a set of meta-features —a form of meta-knowledge. The value of σ for a new dataset is estimated by aggregating the observed best values of σ for the nearest datasets (using a metric over the space of meta-feature values). The output is a ranked sequence of recommended values. The authors show that the method can select a setting with lower error than the default or other comparative approaches.

The second contribution to this issue (Schmidhuber, 2004) is an example of an algorithm that re-learns through experience. It makes use of self-delimiting binary programs to propose an optimal ordered problem solver. The idea is to explore the space of programs that provide a solution to a target problem; new programs are generated either anew, or by re-using previously generated candidate programs. Such search embeds a trade-off between exploration (search for new programs) and exploitation (search for variant solutions). The rationale is that exploiting experience collected in previous search can solve the target problem much faster. Programs are partial solutions to the problem and not composite learners (Section 2.1). Meta-knowledge is stored in the form of candidate program solutions. Exploiting this information is akin to exploiting knowledge for incremental self improvement.

The third contribution to this issue (Dzeroski & Zenko, 2004) is a study within the area of multiple classifiers (Section 2.2), in particular, within the learning paradigm known as stacking. We mentioned before that stacked generalization can be considered a form of meta-learning because the transformation of the training set conveys information about the predictions of the base-learners (i.e., conveys meta-knowledge). In this paper, the authors show the benefits of using new meta-features that capture the confidence on the class posterior probabilities output by a set of base learners. Improved performance is reported when a multi-response linear regression tree is used as the meta-learner. This improvement can be attributed to a change of representation in the original attribute space that appears to simplify the task of the meta-classifier.

The fourth and last contribution to this issue (Kalousis, Gama, & Hilario, 2004) is another example of a learning system comprising multiple classifiers; the problem is to match domain properties with learning performance (Section 2.2). The authors use clustering techniques to characterize relations between datasets and learning algorithms. They find patterns among learners, by clustering the similarity of error correlation distributions of pairs of algorithms across many datasets. In a similar way they find clusters of datasets

having similar patterns of correlations among algorithms. These patterns—a form of meta-knowledge— can bear multiple benefits. One may use them to select a pool of heterogeneous base learners for stacking, or to characterize groups of datasets pointing to a strong dominion of certain learning algorithms, etc. The patterns are derived from an initial characterization of datasets through meta-features (e.g., class entropy, log of the number of examples, ratio of examples to attributes, etc.).

4. Discussion and conclusions

Our goal in this introduction has been to clarify the role meta-knowledge plays in different learning systems. The analysis and understanding of all types of knowledge amenable to extraction from the learning process (i.e. meta-knowledge) is key to the advancement of the field. Under such view, multiple directions are open for future research. For instance, an overview of recent work in meta-learning, including the contributions to this issue, clearly indicates a need for further work in the characterization of datasets (or learning tasks and/or contexts, in general). A proper characterization of datasets is key to the accurate prediction of meta-learning assistants. This holds whether it refers to selecting a good predictive model, estimating model parameters, looking for heterogeneous models, etc.

In addition, further work is necessary to characterize learning algorithms (or strategies in general). Together with a proper characterization of datasets, this would enable us to match learning algorithms with input-output distributions. This implies going beyond a measure of the capacity of the learning machine and its effects on the bias-variance dilemma in statistical inference (Geman, Bienenstock, & Doursat, 1991; Hastie et al., 2001), to a broader understanding of learning strategies and their effect under different dataset characteristics.

Another promising avenue of research lies on exploiting pieces of code in the construction of learning algorithms. Besides the work by Schmidhuber (2004), a similar idea can be pursued using techniques from evolutionary programming. From a meta-learning perspective, however, an interesting approach would be to decompose current learning algorithms to pinpoint specific reasons for their performance according to the example distribution under analysis; we need meta-knowledge indicating how pieces of code can be combined into new learning strategies. Meta-knowledge could be useful to select, combine, or adapt individual constituents to specific tasks, and to reconfigure learning architectures in light of past experience.

Finally, our characterization of learning systems could be expanded to cover complex system made of sets of algorithms employed on different but related tasks. For instance, on a text extraction system, one algorithm may be oriented towards POS tagging, another towards morphosyntactic analysis, yet another towards word sense disambiguation, and so on. The idea is somewhat related to layered learning (Stone & Veloso, 2000; Utgoff & Stracuzzi, 2003), except tasks need not build on top of each other but simply interact with each other. Recent advances in this area indicate that many of those tasks can be acquired through learning. First, meta-knowledge can be used to adapt individual algorithms to specific tasks. In addition, it can be used to control the assignment of algorithms to different tasks, should a choice arise to reconfigure the architecture in light of past experience. In our view, this will prove a critical functionality in future meta-learning tools.

Acknowledgments

We are grateful to Rob Holte, Foster Provost, Karen Cullen, Melissa Fearon, and all the reviewers involved in this special issue for their helpful assistance.

References

Aha, D. W. (1992). Generalizing from case studies: A case study. In *Proceedings of the Ninth International Workshop on Machine Learning* (pp. 1–10), Morgan Kaufman.

Baltes, J., & MacDonald, B. (1992). Case-based meta learning: Sustained learning supported by a dynamically biased version space. In *Proceedings of the ML-92 Workshop on Biases in Inductive Learning*.

Baxter, J. (1998). Theoretical models of learning to learn. *Learning to Learn* (Chap. 4, pp. 71–94), Kluwer Academic Publishers, MA.

Bensusan, H. (1998). God doesn't always shave with Occam's Razor—Learning when and how to prune. In *Proceedings of the Tenth European Conference on Machine Learning* (pp. 119–124), Springer.

Bensusan, H., & Giraud-Carrier, C. (2000). Casa Batlo in Passeig or landmarking the expertise space. In *Proceedings of the ECML-2000 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination* (pp. 29–46), Barcelona, Spain.

Brazdil, P., Soares, C., & Pinto da Costa, J. (2003). Ranking learning algorithms: Using IBL and meta-learning on accuracy and time results. *Machine Learning, 50:3*, 251–277.

Brazdil, P. (1998). Data transformation and model selection by experimentation and meta-learning. In *Proceedings of the ECML-98 Workshop on Upgrading Learning to Meta-Level: Model Selection and Data Transformation* (pp. 11–17), Technical University of Chemnitz.

Brodley, C. E. (1995). Recursive automatic bias selection for classifier construction. *Machine Learning, 20:1*, 63–94.

Caruana, R. (1997). Multitask Learning. *Second Special Issue on Inductive Transfer. Machine Learning, 28:1*, 41–75.

Chan, P. K., & Stolfo, S. (1998). On the accuracy of meta-learning for scalable data mining. *Journal of Intelligent Integration of Information 8*, 3–28.

DesJardins, M., & Gordon, D. F. (1995). Evaluation and selection of biases in machine learning. *Machine Learning, 20:1*, 5–22.

Dzeroski, S., & Zenko, B. (2004). Is combining classifiers better than selecting the best one? *Machine Learning, 54:3*, 195–209.

Gama, J., & Brazdil, P. (1995). Characterization of classification algorithms. In *Proceedings of the Seventh Portuguese Conference on Artificial Intelligence (EPIA)* (189–200), Funchal, Madeira Island, Portugal.

Geman, S., Bienenstock, E., & Doursat, R. (1991). Neural networks and the bias/variance dilemma. *Neural Computation, 4*, 1–58.

Gordon, D. F. (1990). Active bias adjustment for incremental, supervised concept learning. PhD Thesis, University of Maryland, 1990.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning; Data Mining, Inference, and Prediction*. Springer-Verlag.

Kalousis, A., Gama, J., & Hilario, M. (2004). On data and algoritms: Understanding learning performance. *Machine Learning, 54:3*, 195–209.

Keller, J., Paterson, I., & Berrer, H. (2000). An integrated concept for multi-crieria-ranking of data-mining algorithms. In *Proceedings of the ECML-2000 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination* (pp. 73–86), Barcelona, Spain.

Merz, C. J. (1995). Dynamical selection of learning algorithms. In *Learning from Data: Artificial Intelligence and Statistics*, D. Fisher & H. J. Lenz (Eds.), Springer-Verlag.

Michie, D., Spiegelhalter, D. J., & Taylor, C. C. (1994). *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Chichester, England.

Ortega, J., Koppel, M., & Argamon, S. (2001). Arbitrating among competing classifiers using learned referees. *Knowledge and Information Systems, 3*, 470–490.

Peng, W., Flach, P. A., Soares, C., & Brazdil, P. (2002). Improved Data Set Characterisation for Meta-learning. In *Proceedings of the Fifth International Conference on Discovery Science, LNAI 2534*, 141–152.

Pfahringer, B., Bensusan, H., & Giraud-Carrier, C. (2000). Meta-learning by landmarking various learning algorithms. In *Proceedings of the Seventeenth International Conference on Machine Learning* (743–750), Stanford, CA.

Pratt, L., & Thrun, S. (1997). Second Special Issue on Inductive Transfer. *Machine Learning*, 28(1).

Rendell, L., Seshu, R., & Tcheng, D. (1987a). More robust concept learning using dynamically-variable bias. In *Proceedings of the Fourth International Workshop on Machine Learning* (pp. 66–78), Morgan Kaufman.

Rendell, L., Seshu, R., & Tcheng, D. (1987b). Layered concept-learning and dynamically-variable bias management. In *Proceedings of the International Joint Conference of Artificial Intelligence*, (pp. 308–314), Milan, Italy.

Schmidhuber, J. (2004). Optimal ordered problem solver. *Machine Learning, 54:3*, 195–209.

Soares, C., Brazdil, P., & Kuba, P. (2004). A meta-learning approach to select the kernel width in support vector regression, *Machine Learning, 54:3*, 195–209.

Stone, P., & Veloso, M. (2000). Layered learning, In *Proceedings of the Eleventh European Conference on Machine Learning* (pp. 369–381) Barcelona, Spain.

Todorovski, L., & Dzeroski, S. (2003). Combining classifiers with meta decision trees *Machine Learning, 50:3*, 223–250.

Thrun, S., & Mitchell, T. (1995). Learning one more thing. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 1217–1223), Morgan Kaufman.

Thrun, S. (1998). Lifelong learning algorithms. *Learning to Learn* (Chap. 8, pp. 181–209), MA: Kluwer Academic Publishers.

Utgoff, P. (1986). Shift of bias for inductive concept learning. In R. S. Michalski, et al. (Ed.), *Machine Learning: An Artificial Intelligence Approach*, Vol. II (pp. 107–148), California: Morgan Kaufman.

Utgoff, P., & Stracuzzi, D. J. (2003). Many-layered learning. *Neural Networks, 14*, 2497–2529, MIT Press.

Vilalta, R., & Drissi, Y. (2002). A perspective view and survey of meta-learning. *Journal of Artificial Intelligence Review, 18:2*, 77–95.

Widmer, G., & Kubat, M. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23:1*, 69–101.

Widmer, G. (1997). Tracking context changes through meta-learning. *Machine Learning, 27:3*, 259–286.

Wolpert, D. (1992). Stacked Generalization. *Neural Networks, 5*, 241–259.