Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume III
K–Pri

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey • New York

Director of Editorial Content:	Kristin Klinger
Director of Production:	Jennifer Neidig
Managing Editor:	Jamie Snavely
Assistant Managing Editor:	Carole Coulson
Typesetter:	Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design:	Lisa Tosheff
Printed at:	Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

All work contributed to this encyclopedia set is new, previously-unpublished material. The views expressed in this encyclopedia set are those of the authors, but not necessarily of the publisher.

If a library purchased a print copy of this publication, please go to http://www.igi-global.com/agreement for information on activating the library's complimentary electronic access to this publication.

Meta-Learning

Christophe Giraud-Carrier
Brigham Young University, USA

Pavel Brazdil
University of Porto, Portugal

Carlos Soares
University of Porto, Portugal

Ricardo Vilalta
University of Houston, USA

INTRODUCTION

The application of Machine Learning (ML) and Data Mining (DM) tools to classification and regression tasks has become a standard, not only in research but also in administrative agencies, commerce and industry (e.g., finance, medicine, engineering). Unfortunately, due in part to the number of available techniques and the overall complexity of the process, users facing a new data mining task must generally either resort to trial-and-error or consultation of experts. Clearly, neither solution is completely satisfactory for the non-expert end-users who wish to access the technology more directly and cost-effectively.

What is needed is an informed search process to reduce the amount of experimentation with different techniques while avoiding the pitfalls of local optima that may result from low quality models. Informed search requires meta-knowledge, that is, knowledge about the performance of those techniques. Meta-learning provides a robust, automatic mechanism for building such meta-knowledge. One of the underlying goals of meta-learning is to understand the interaction between the mechanism of learning and the concrete contexts in which that mechanism is applicable. Meta-learning differs from base-level learning in the scope of adaptation. Whereas learning at the base-level focuses on accumulating experience on a specific learning task (e.g., credit rating, medical diagnosis, mine-rock discrimination, fraud detection, etc.), learning at the meta-level is concerned with accumulating experience on the performance of multiple applications of a learning system.

The meta-knowledge induced by meta-learning provides the means to inform decisions about the precise conditions under which a given algorithm, or sequence of algorithms, is better than others for a given task. While Data Mining software packages (e.g., SAS Enterprise Miner, SPSS Clementine, Insightful Miner, PolyAnalyst, KnowledgeStudio, Weka, Yale, Xelopes) provide user-friendly access to rich collections of algorithms, they generally offer no real decision support to non-expert end-users. Similarly, tools with emphasis on advanced visualization help users understand the data (e.g., to select adequate transformations) and the models (e.g., to tweak parameters, compare results, and focus on specific parts of the model), but treat algorithm selection as a post-processing activity driven by the users rather than the system. Data mining practitioners need systems that guide them by producing explicit advice automatically. This chapter shows how meta-learning can be leveraged to provide such advice in the context of algorithm selection.

BACKGROUND

STABB is an early precursor of meta-learning systems in the sense that it was the first to show that a learner's bias can be adjusted dynamically (Utgoff, 1986). VBMS may be viewed as the first simple meta-learning system (Rendell et al., 1989). It learns to choose one of three symbolic learning algorithms as a function of the number of training instances and the number of features. The StatLog project extended VBMS significantly by considering a larger number of dataset characteristics, together with a broad class of candidate models and algorithms for selection (Brazdil & Henery, 1994). The aim was to characterize the space in which a given algorithm achieves positive generalization performance.

The MLT project focused on the practice of machine learning and produced a toolbox consisting of a number of learning algorithms, datasets, standards and know-how (Kodratoff et al., 1992; Craw et al., 1992). Considerable insight into many important machine learning issues was gained during the project, much of which was translated into meta-rules that formed the basis of a kind of user-guidance expert system called Consultant-2.

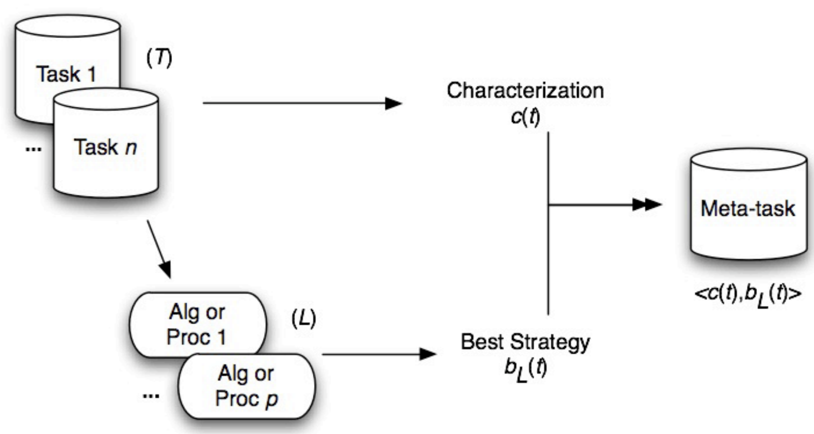
Born out of practical challenges faced by researchers at Daimler Benz AG (now), CITRUS is perhaps the first implemented system to offer user guidance for the complete data mining process, rather than for a single phase of the process (Engels, 1996; Wirth et al., 1997). Algorithm selection takes place in two stages, consisting of: 1) mapping tasks to classes of algorithms, and 2) selecting an algorithm from the selected class. The mapping stage is driven by decomposition and guided by high-level pre/post-conditions (e.g., interpretability). The selection stage consists of using data characteristics (inspired by the Statlog project) together with a process of elimination (called "strike-through"), where algorithms that do not work for the task at hand are successively eliminated until the system finds one applicable algorithm. Although there is no meta-learning in the traditional sense in CITRUS, there is still automatic guidance beyond the user's own input.

Finally, theoretical results, such as the NFL theorems and their consequences have helped in identifying limitations and opportunities for meta-learning (Schaffer, 1994; Wolpert & Macready, 1995; Wolpert, 2001). Additionally, extensive empirical studies have confirmed the theory, and provided additional insight into learning that may serve both as a source of direct meta-knowledge and as input to meta-learning (Aha, 1992; Holte, 1993; Lim et al., 2000).[1]

MAIN FOCUS

Meta-learning, in the context of model selection, consists of applying learning mechanisms to the problem of mapping learning tasks to algorithms. Let L be a set of learning algorithms and T be a set of learning tasks such that for each t in T, $b_L(t)$ represents the algorithm in L that performs best on t for some user-defined performance criterion (e.g., predictive accuracy, execution time).[2] Since learning tasks may be unwieldy to handle

Figure 1. Meta-dataset construction



directly, some type of task characterization is used and the meta-learner actually learns a mapping from characterizations to algorithms. For each learning task t, let $c(t)$ denote the characterization of t by some fixed mechanism. The set $\{<c(t), b_L(t)>: t \text{ in } T\}$ constitutes a meta-task or meta-dataset, as depicted in Figure 1.

A meta-learner can then take the meta-dataset $\{<c(t), b_L(t)>: t \text{ in } T\}$ as a training set and induce a meta-model that, for each new learning task, predicts the model from L that will perform best. Alternatively, one may build a meta-model that predicts a ranking of algorithms from L (Berrer et al., 2000; Brazdil et al., 2003). The ranking approach reduces the brittleness of the meta-model. Assume, for example, that the model predicted best for some new learning task results in what appears to be a poor performance. In the single-model prediction approach, the user has no further information as to what other model to try. In the ranking approach, the user may try the second best, third best, and so on, in an attempt to improve performance. Furthermore, ranking makes it easier to include additional (possibly qualitative) criteria, such as comprehensibility, in the selection process (Giraud-Carrier, 1998).

Clearly, one of the challenges of meta-learning is the construction of the meta-dataset, i.e, $<c(t), b_L(t)>$ pairs for some base level learning tasks. This raises issues with: 1) the choice of the characterization mechanism c, 2) the choice of the set of learners L, 3) the collection of representative tasks, and 4) the cost of computing $c(t)$ and $b_L(t)$ for each task. We briefly discuss each of these issues in the following sections.

Characterization Mechanism

As in any learning task, the characterization of the examples plays a crucial role in enabling learning. The central idea is that high-quality dataset characteristics or meta-features provide useful information to discriminate among the performances of a set of given learning strategies. Typical characterization techniques belong to one of the following classes.

- *Statistical and Information-Theoretic Characterization.* A number of statistical and information-theoretic measures are extracted from the dataset, such as number of classes, number of features, ratio of examples to features, degree of correlation between features and target concept, average class entropy and class-conditional entropy, skewness, kurtosis, signal–to-noise ratio, etc. (Aha, 1992; Michie et al., 1994; Engels & Theusinger, 1998; Sohn, 1999; Köpf et al., 2000; Kalousis, 2002).
- *Model-Based Characterization*. Models induced on the dataset are used as indicators of the underlying properties of the dataset. To date, only decision trees have been used for the extraction of characteristics such as nodes per feature, maximum tree depth, shape, tree imbalance, etc. (Bensusan et al., 2000; Peng et al., 2002).
- *Landmarking*. The performances of simple learners, known as landmarkers, are computed on the dataset using cross-validation (Bensusan & Giraud-Carrier, 2000; Pfahringer et al., 2000). The idea is that landmarkers serve as signposts of the performance of the full-fledged target learners in L. Alternatively, one can exploit accuracy results obtained on simplified versions of the data (e.g., samples), known as sub-sampling landmarks (Fürnkranz & Petrak, 2001; Soares et al., 2001).

Choice of Base-level Learners

Although no learner is universal, each learner has its own area of expertise, which can be informally defined as the set of learning tasks on which it performs well. Since the role of the meta-model is to predict which algorithm is most likely to perform best on each new task, one should select base learners with complementary areas of expertise. In principle, one should seek the smallest set of learners that is most likely to ensure a reasonable coverage of the base-level learning space. One way to ensure diversity is by choosing representative learners from varied model classes. The more varied the biases, the greater the coverage.

Meta-Dataset Construction

The number of accessible, documented, real-world learning tasks is relatively small, which poses a challenge for learning. This challenge may be addressed either by augmenting the meta-dataset through systematic generation of synthetic base level tasks, or by taking the view that the model selection task is inherently incremental and treating it as such. The second approach results in slower learning since learning tasks become available over time. On the other hand, it naturally adapts to reality, extending to new areas of the base level learning space only when tasks from these areas actually arise.

Computational Cost

The computational cost is the price to pay to be able to perform model selection learning at the meta-level. However, in order to be justifiable, the cost of computing $c(t)$ should be significantly lower than the cost of computing $b_L(t)$. Otherwise, even if the meta-model is very accurate, it has little value as the user would be better off trying all algorithms and selecting the best one, which clearly defeats the purpose of meta-learning. The characterization mechanisms listed above all include many measures that satisfy this condition.

Although much remains to be done, results suggest the suitability of meta-learning for model selection (Brazdil & Soares, 2000; Bensusan & Kalousis, 2001; Hilario & Kalousis, 2001). We briefly describe two recent, successful systems as an illustration. One is a strict meta-learning system and offers ranking advice for model selection. The other is based on an ontology, but produces ranking advice for the complete KDD process.

Data Mining Advisor

The Data Mining Advisor (DMA) is the main product of the ESPRIT METAL research project (see http://www.metal-kdd.org). The DMA is a Web-based meta-learning system for the automatic selection of model building algorithms in the context of classification and regression tasks. Given a dataset and goals defined by the user in terms of accuracy and training time, the DMA returns a list of algorithms that are ranked according to how well they are predicted to meet the stated goals.

The DMA guides the user through a wizard-like step-by-step process consisting of 1) uploading a target dataset (with some user-defined level of privacy), 2) characterizing the dataset automatically using statistical and information-theoretic measures, 3) setting the selection criteria and the ranking method, 4) producing the ranking advice, and 5) executing user-selected algorithms on the dataset. Although the induced models themselves are not returned, the DMA reports 10-fold cross-validation accuracy, true rank and score, and, when relevant, training time. A simple example of the ranking produced by the DMA is shown in Figure 2, where some algorithms were selected for execution.

The DMA's choice of providing rankings rather than "best-in-class" is motivated by a desire to give as much information as possible to the user, as discussed above. In some sense, one can argue that the ranking approach subsumes the "best-in-class" approach. Interestingly, empirical evidence suggests that the best algorithm is generally within the top 3 in the DMA rankings (Brazdil et al., 2003).

Intelligent Discovery Assistant

The notion of Intelligent Discovery Assistant (IDA) provides a template for building ontology-driven, pro-

Figure 2. Sample DMA output

Ranking table
Dowload results | register now! | learn more about the online advisor

Predicted Rank	Algorithm	Predicted Score	Status	Run	Accuracy	Time	True Rank	True Score
1.	c50rules	1.031	finished	--	0.2830000	?	6	1.003
2.	lindiscr	1.03	finished	--	0.2340000	?	1	1.048
3.	c50tree	1.026	--		--	--	--	--
4.	ltree	1.023	--		--	--	--	--
5.	clemMLP	1.017	finished	--	0.2680000	?	5	1.006
6.	c50boost	1.017	--		--	--	--	--
7.	ripper	1.009	--		--	--	--	--
8.	mlcnb	1	finished	--	0.2430000	?	2	1.036
9.	clemRBFN	0.948	--		--	--	--	--
10.	mlcib1	0.913	finished	--	0.3180000	?	10	0.938

cess-oriented assistants for KDD (Bernstein & Provost, 2001; Bernstein et al., 2005). IDAs encompass the three main algorithmic steps of the KDD process, namely, pre-processing, model building and post-processing. In IDAs, any chain of operations consisting of one or more operations from each of these steps is called a Data Mining (DM) process. The goal of an IDA is to propose to the user a list of ranked DM processes that are both valid and congruent with user-defined preferences (e.g., speed, accuracy).

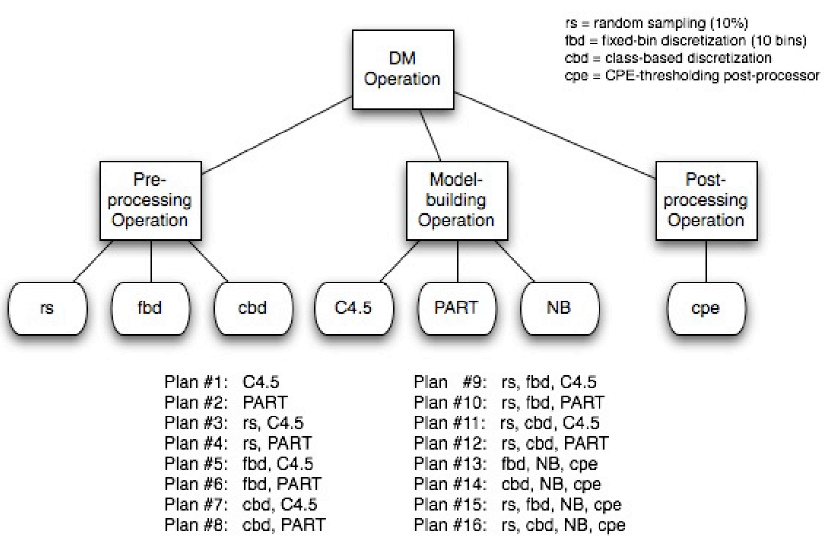
The IDA's underlying ontology is essentially a taxonomy of DM operations or algorithms, where the leaves represent implementations available in the corresponding IDA. Operations are characterized by pre-conditions, post-conditions and heuristic indicators. Clearly, the versatility of an IDA is a direct consequence of the richness of its ontology. The typical organization of an IDA consists of 1) the plan generator, that uses the ontology to build a list of (all) valid DM processes that are appropriate for the task at hand, and 2) the heuristic ranker, that orders the generated DM processes according to preferences defined by the user.

The plan generator takes as input a dataset, a user-defined objective (e.g., build a fast, comprehensible classifier) and user-supplied information about the data, i.e., information that may not be obtained automatically. Starting with an empty process, it systematically searches for an operation whose pre-conditions are met and whose indicators are congruent with the user-defined preferences. Once an operation has been found, it is added to the current process, and its post-conditions become the system's new conditions from which the search resumes. The search ends once a goal state has been found or when it is clear that no satisfactory goal state may be reached. The plan generator's search is exhaustive: all valid DM processes are computed. Figure 3 shows the output of the plan generator for a small ontology of only 7 operations, when the input dataset is continuous-valued and comprehensible classifiers are to be preferred.

The restriction of the plan generator to valid processes congruent with user-defined objectives is generally sufficient to make an exhaustive search feasible. The main advantage of this exhaustivity is that no valid DM process is ever overlooked, as is likely to be the case with most users, including experts. As a result, an IDA may, and evidence suggests that it does, uncover novel processes that experts had never thought about before, thus enriching the community's meta-knowledge (Bernstein & Provost, 2001).

Once all valid DM processes have been generated, a heuristic ranker is applied to assist the user further, by organizing processes in descending order of ``return'' on user-specified goals. For example, the processes in Figure 3 are ordered from simplest (i.e., least number of steps) to most elaborate. The ranking relies on the knowledge-based heuristic indicators. If speed rather than simplicity were the objective, for example, then

Figure 3. Sample list of IDA-generated DM processes



Plan #3 would be bumped to the top of the list, and all plans involving random sampling would also move up. In the current implementation of IDAs, rankings rely on fixed heuristic mechanisms. However, IDAs are independent of the ranking method and thus, they could possibly be improved by incorporating meta-learning to generate rankings based on past performance.

FUTURE TRENDS

One important research direction in meta-learning consists in searching for improved meta-features in the characterization of datasets. A proper characterization of datasets can elucidate the interaction between the learning mechanism and the task under analysis. Current work has only started to unveil relevant meta-features; clearly much work lies ahead. For example, many statistical and information-theoretic measures adopt a global view of the dataset under analysis; meta-features are obtained by averaging results over the entire training set, implicitly smoothing the actual distribution. There is a need for alternative and more detailed descriptors of the example distribution in a form that highlights the relationship to the learner's performance.

Similarly, further research is needed in characterizing learning algorithms. Recent efforts in model composition may prove useful. In this paradigm, instead of seeking to combine several whole algorithms or to find one algorithm among several that would perform best on a given task, the system breaks the learning process down into sub-components and, for each task, composes a custom, possibly novel, learning system from a combination of these components (Suyama et al., 1998; Abe & Yamaguchi, 2002).

Recently proposed agent-based data mining architectures offer unique ways to increase the versatility, extensibility and usability of meta-learning (Botía et al., 2001; Hernansaez et al., 2004; Zhong et al., 2001).

Finally, the increased amount and detail of data available about the operations of organizations is leading to a demand for a much larger number of models, up to hundreds or even thousands. This kind of application has been called Extreme Data Mining (Fogelman-Soulié, 2006). Current DM methodologies, which are largely dependent on human efforts, are not suitable for this kind of extreme settings because of the large amount of human resources required. Meta-learning can be used to reduce the need for human intervention in model development and thus, we expect that it will play a major role in these large-scale Data Mining applications.

CONCLUSION

From a practical standpoint, meta-learning helps solve important problems in the application of data mining tools. First, the successful use of these tools outside the boundaries of research is conditioned upon the appropriate selection of a suitable predictive model, or combination of models, according to the domain of application. Without some kind of assistance, model selection and combination can turn into solid obstacles to non-expert end-users. Second, a problem commonly observed in the practical use of data mining tools is how to profit from the repetitive use of predictive models over similar tasks. The successful application of models in real-world scenarios requires continuous adaptation to new needs. Rather than starting afresh on new tasks, one would expect the learning mechanism itself to re-learn, taking into account previous experience. Again, meta-learning systems can help control the process of exploiting cumulative expertise by searching for patterns across tasks, thus improving the utility of data mining. Interestingly, generalizations of meta-learning for algorithm selection in other areas, such as cryptography, sorting and optimization, have recently been proposed (Smith-Miles, 2007).

REFERENCES

Abe, H., & Yamaguchi, T. (2002). Constructing Inductive Applications by Meta-Learning with Method Repositories, in *Progress in Discovery Science, Final Report of the Japanese Discovery Science Project*, LNCS 2281, Springer-Verlag, 576-585.

Aha D. W. (1992). Generalizing from Case Studies: A Case Study, in *Proceedings of the Ninth International Workshop on Machine Learning*, 1-10.

Bensusan, H., & Giraud-Carrier, C. (2000). Discovering Task Neighbourhoods Through Landmark Learning Performances, in *Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases*, 325,330.

Bensusan H., Giraud-Carrier C., & Kennedy C.J. (2000). A Higher-Order Approach to Meta-Learning, in *Proceedings of the ECML-2000 Workshop on Meta-Learning: Building Automatic Advice Strategies for Model Selection and Method Combination,* 109-118.

Bensusan, H., & Kalousis, A. (2001). Estimating the Predictive Accuracy of a Classifier, in *Proceedings of the Twelfth European Conference on Machine Learning*, 25-36.

Bernstein, A., & Provost, F. (2001). An Intelligent Assistant for the Knowledge Discovery Process, in *Proceedings of the IJCAI-01 Workshop on Wrappers for Performance Enhancement in KDD*.

Bernstein, A., Provost, F., & Hill, S. (2005). Toward Intelligent Assistance for a Data Mining Process: An Ontology-based Approach for Cost-sensitive Classification, *IEEE Transactions on Knowledge and Data Engineering*, 17(4):503-518.

Berrer, H., Paterson, I., & Keller, J. (2000). Evaluation of Machine-learning Algorithm Ranking Advisors, in *Proceedings of the PKDD-2000 Workshop on Data-Mining, Decision Support, Meta-Learning and ILP: Forum for Practical Problem Presentation and Prospective Solutions*.

Botía, J.A., Gómez-Skarmeta, A.F., Valdés, M., & Padilla, A. (2001). METALA: A Meta-learning Architecture, in *Proceedings of the International Conference, Seventh Fuzzy Days on Computational Intelligence, Theory and Applications* (LNCS 2206), 688-698.

Brazdil, P., & Henery, B. (1994). Analysis of Results, in Michie, D. et al. (Eds.), *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, Chapter 10.

Brazdil, P., & Soares, C. (2000). A Comparison of Ranking Methods for Classification Algorithm Selection, in *Proceedings of the Eleventh European Conference on Machine Learning*, 63-74.

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.

Craw, S., Sleeman, D., Granger, N., Rissakis, M., & Sharma, S. (1992). Consultant: Providing Advice for the Machine Learning Toolbox, in Bramer, M., & Milne, R. (Eds.), *Research and Development in Expert Systems IX (Proceedings of Expert Systems '92)*, SGES Publications, 5-23.

Engels, R. (1996). Planning Tasks for Knowledge Discovery in Databases; Performing Task-Oriented User-Guidance, in *Proceedings of the Second ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 170-175.

Engels, R., & Theusinger, C. (1998). Using a Data Metric for Offering Preprocessing Advice in Data-mining Applications, in *Proceedings of the Thirteenth European Conference on Artificial Intelligence*.

Fogelman-Soulié, F. (2006). Data Mining in the Real World: What Do We Need and What Do We Have? in *Proceedings of the KDD-2206 Workshop on Data Mining for Business Applications*, 44-48.

Fürnkranz, J., & Petrak J. (2001). An Evaluation of Landmarking Variants, in *Working Notes of the ECML/PKDD 2001 Workshop on Integrating Aspects of Data Mining, Decision Support and Meta-Learning*.

Giraud-Carrier, C. (1998). Beyond Predictive Accuracy: What?, in *Proceedings of the ECML-98 Workshop on Upgrading Learning to the Meta-Level: Model Selection and Data Transformation,* 78-85.

Hernansaez, J.M., Botía, J.A., & Gómez-Skarmeta, A.F. (2004). A J2EE Technology Based Distributed Software Architecture for Web Usage Mining, in *Proceedings of the Fifth International Conference on Internet Computing*, 97-101.

Hilario, M., & Kalousis, A. (2001). Fusion of Meta-Knowledge and Meta-Data for Case-Based Model Selection, in *Proceedings of the Fifth European Conference on Principles of Data Mining and Knowledge Discovery*, 180-191.

Holte, R.C. (1993). Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, *Machine Learning*, 11:63-91.

Kalousis, A. (2002). Algorithm Selection via Meta-Learning, Ph.D. Thesis, University of Geneva, Department of Computer Science.

Kodratoff, Y., Sleeman, D., Uszynski, M., Causse, K., & Craw, S. (1992). Building a Machine Learning Toolbox, in Steels, L., & Lepape, B. (Eds.), *Enhancing the Knowledge Engineering Process*, Elsevier Science Publishers, 81-108.

Köpf, C., Taylor, C.C., & Keller, J. (2000). Meta-analysis: From Data Characterisation for Meta-learning to Meta-regression, in *Proceedings of the PKDD-2000 Workshop on Data Mining, Decision Support, Meta-learning and ILP: Forum for Practical Problem Presentation and Prospective Solutions*.

Lim, T-S., Loh, W-Y., & Shih, Y-S. (2000). A Comparison of Prediction Accuracy, Complexity and Training Time of Thirty-Three Old and New Classification Algorithms, *Machine Learning*, 40:203-228.

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

Peng, Y., Flach, P., Brazdil, P., & Soares, C. (2002). Decision Tree-Based Characterization for Meta-Learning, in *Proceedings of the ECML/PKDD'02 Workshop on Integration and Collaboration Aspects of Data Mining, Decision Support and Meta-Learning,* 111-122.

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.

Rendell, L., Seshu, R., & Tcheng, D. (1989). Layered Concept Learning and Dynamical Variable Bias Management, in *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*.

Schaffer, C. (1994). A Conservation Law for Generalization Performance, in *Proceedings of the Eleventh International Conference on Machine Learning*, 259-265.

Smith-Miles, K. (2007). Cross-disciplinary Perspectives on the Algorithm Selection Problem: Recent Advances and Emerging Opportunities, Technical Report TR C07/07, Deakin University, School of Engineering and Information Technology.

Soares, C., Petrak, J., & Brazdil, P. (2001). Sampling-Based Relative Landmarks: Systematically Test-Driving Algorithms before Choosing, in *Proceedings of the Tenth Portuguese Conference on Artificial Intelligence*.

Sohn, S.Y. (1999). Meta Analysis of Classification Algorithms for Pattern Recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(11):1137-1144.

Suyama, A., Negishi, N., & Yamaguchi, T. (1998). CAMLET: A Platform for Automatic Composition of Inductive Learning Systems Using Ontologies, in *Proceedings of the Fifth Pacific Rim International Conference on Artificial Intelligence*, 205-215.

Utgoff, P.E. (1986). Shift of Bias for Inductive Concept Learning, in Michalski, R.S., Carbonell, J.G., & Mitchell, T.M. (Eds.), *Machine Learning: An Artificial Intelligence Approach, Volume II*, Morgan Kaufmann Publishers. Inc., Chapter 5.

van Someren, M. (2001). Model Class Selection and Construction: Beyond the Procustean Approach to Machine Learning Applications, in Paliouras, G., Karkaletsis, V., & Spyropoulos, C.D. (Eds.), *Machine Learning and Its Applications: Advanced Lectures, LNCS 2049*, Springer-Verlag, 196-217.

Wirth, R., Shearer, C., Grimmer, U., Reinartz, T.P., Schlosser, J., Breitner, C., Engels, R., & Lindner, G. (1997). Towards Process-Oriented Tool Support for Knowledge Discovery in Databases, in Proceedings of the First European Conference on Principles and Practice of Knowledge Discovery in Databases, 243-253.

Wolpert, D.H., & Macready, W.G. (1995). No Free Lunch Theorems for Search, Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Wolpert, D.H. (2001). The Supervised Learning No-Free-Lunch Theorems, in *Proceedings of the Sixth Online World Conference on Soft Computing in Industrial Applications*, 325-330.

Zhong, N., Liu, C., & Oshuga, S. (2001). Dynamically Organizing KDD Processes, *International Journal of Pattern Recognition and Artificial Intelligence*, 15(3):451-473.

KEY TERMS

Data Mining: Application of visualization, statistics and machine learning to the discovery of patterns in databases. There is general consensus that patterns found by data mining should in some way be novel and actionable.

Landmarking: A task characterization that replaces a learning task by the performances of a number of simple and efficient learning algorithms on that task.

Meta-Dataset: Dataset consisting of task characterizations (or meta-features) together with their associated best strategy (i.e., learning algorithm or data mining process that gives the best performance on the task).

Meta-Features: Features used to characterize datasets, that serve as inputs to meta-learning. These features may take the form of statistics, landmarkers or model-based attributes.

Meta-Learning: Application of learning techniques at the meta-level. Any use of learning methods to help inform the process of machine learning. Learning about learning.

Task Characterization: A method for extracting features, then known as meta-features, from the dataset associated with a learning task.

ENDNOTES

[1] Note that, although it is sometimes viewed as a form of meta-learning, we purposely omit from this discussion the notion of model combination. Model combination consists of creating a single learning system from a collection of learning algorithms. It has been shown that in many cases improved performance is obtained by combining the strengths of several learning algorithms. These approaches reduce the probability of misclassification based on any single induced model by increasing the system's area of expertise through combination. However, from the meta-learning perspective, they can be regarded as single algorithms.

[2] It is easy to extend the notion of best learning algorithm for t to best data mining process for t.