

# Preface

Miscellaneous learning machines have been designed and implemented for decades of computational intelligence (CI) research. More and more solutions are being published all the time in all the subdomains like classification, regression, clustering, and others. A natural question arises: how many algorithms to solve the same goals must be proposed, before we regard the available set of tools as satisfactory? One could claim that such time will never come, because there will always be a possibility of improvement in various aspects, various contexts of applications. This is certainly true, but on the other side, it seems quite reasonable to claim something completely opposite: we already have so many tools that, to solve new tasks, we no longer need new methods, but a robust knowledge about how to use the tools we have already got. Probably, each real-world task fitting the domain of computational intelligence can be solved by available tools in a way close to optimal (statistically insignificantly different than optimal), but it does not mean that finding a satisfactory solution is easy. Therefore, more effort of the CI community should be focused on meta-level algorithms, capable of finding attractive solutions with already available tools, than on development of new learning methods that can solve some new, special types of problems.

This book is focused on learning decision tree (DT) models. One of its goals is to show that even the part of CI devoted to DT induction is large enough to make the pursuit of the most successful learning machine a very complex problem. To be successful in assigning adequate learning machines for particular learning problems, we need more systematic research in the field. The number of available solutions is even larger than we usually think, because each learning machine consists of many components and each of them can be easily replaced by many other compatible modules. Only well designed, general architectures of learning machines reveal the real level of complexity of the problem, because in such universal and flexible frameworks, available components may be combined in huge numbers of ways. No human expert can try all these combinations when solving a problem, so even the best experts are likely to miss quite simple and attractive solutions. Therefore, we need automated tools performing reliable search processes in the space of possible learning machines. Because all possible machines cannot be tested, we need trustworthy metaknowledge helpful in recognizing the more and the less interesting algorithms in various contexts. Starting

with the metaknowledge expressed by human experts and gathering new knowledge, extracted from extensive, automated search processes, may easily bring large knowledge bases of great value for automated expert systems. The metaknowledge may be used not only to select probably the most accurate learning machines, but also to construct new complex learning machines. The amount of such knowledge is so large that very soon, no human expert will be able to use it in an optimal way. Therefore, in many fields, the era of human designed models is close to its end—automated learners will soon outperform humans as they already do in playing chess, because in data modeling, also so large databases of metaknowledge must be searched through and analyzed in appropriate order, that it gets unfeasible for humans, but more and more eligible for automated learners.

The goal of the book is to propose some steps in the direction that seems the most adequate: toward easy to explore taxonomy of DT induction methods and automated tools for construction of successful DT learning machines.

Efficient exploitation of available algorithms and even construction of new algorithms are possible only when the nature of the methods is perfectly understood. Therefore, before writing the book got started, plenty of algorithms had been analyzed, redesigned, implemented, and validated. It is not easy to implement faithfully even the most popular algorithms, because some detailed solutions are often kept hidden, but sometimes they are quite significant. Thus, apart from the analysis of available descriptions (not always clear and exhaustive), sometimes, the source codes had to be analyzed to discover some detailed solutions. In some other cases, when the source codes were not available, some elements of algorithm definition had to be guessed or inferred from the analysis of outputs generated by the binary codes.

[Chapter 2](#) is the result of the work in this area done for many years. It presents many algorithms (more and less popular), described from the point of view of a scientist trying to discover all important aspects of the solutions. Therefore, it is not just a review of selected methods, but a survey of the domain with thorough analysis of advantages and drawbacks, possibilities and limitations of many ideas applied to DT induction. The goal of this chapter was to describe the most important solutions shortly but quite exhaustively, intuitively, in common language, but also with formal statements when necessary to make the algorithms unambiguous. This should allow the readers easily understand the algorithms in relatively short time.

After the survey of DT induction research, the book presents a unified view of the algorithms ([Chap. 3](#)) that facilitates easy combination of many compatible components into new learning machines. Such a framework is absolutely necessary for advanced meta-learning purposes.

Advanced metalearning requires a robust environment for efficient running machine processes, conducting complex tests, collecting, and analysis of the results. [Chapter 4](#) presents the most important solutions of Intemi system, designed and implemented especially for the purposes of metalearning. It arose from many years of experience with learning machines and very deep analysis of meta-learning requirements.

Some aspects of meta-level analysis of DT components in action are presented in [Chap. 5](#), where the problems of reliable testing are discussed, some experiments designed, and their results collected and compared. Adequate methods of result visualization have been worked out to facilitate reliable conclusions. Appropriate visualization of results is very important for humans, but gets quite difficult when tens of thousands of results need to be presented together.

The coping stone of the work described in the book is the meta-learning approach based on search and validation, presented in [Chap. 6](#) with two particular implementations of the idea: One based on machine configuration generators and complexity control, and one dealing with learning results profiles.

[Chapter 7](#) discusses the possible future of meta-learning research and points the directions that seem the most promising and most adequate to follow.

Two appendices contain some descriptions of basic statistical and algebraic methods often used in DT induction. They have been estimated as useful for each inquisitive novice in the field, who may want to get more details without searching outside of the book.

The contents of the book have been prepared thanks to many years of research on DT induction, on learning machines in general, and recently on metalearning. Some fragments of the book review the work of numerous authors, published in miscellaneous articles. The original papers are always pointed as the sources of the information. Some other fragments concern ideas presented by the book author in some publications, but here they are displayed in the context of this monograph—subordinate to the goal of the book, which is presenting the long and complex road from various small and larger algorithms, to a unified approach and the robustness of metalearning.

The book is addressed both to experienced machine learning scientists, interested in the research on DT induction, and to newcomers to the field. A novice should find the review parts ([Chap. 2](#) and [Sect. 6.1](#)) especially useful, as they can help to understand many popular methods of DT induction and metalearning. Most experts shall be more attracted by the final chapters, which take up the topic of metalearning, however also many details of DT induction algorithms are likely to interest even the experienced researchers, because the information presented there is a result of a thorough analysis of the methods, performed by a researcher aiming at understanding various computational aspects and practical advantages of the algorithms. Moreover, the review parts may be seen as an interesting repository of knowledge about many algorithms, presented in a succinct but usually exhaustive descriptions, thus very valuable also for experts. The review and analysis of DT induction approaches, presented in the book, is probably the most extensive one and the most in-depth study, published so far.

Both novices to the field and experts may be interested in the unified view of DT induction algorithms ([Chap. 3](#)) and the general machine learning framework architecture ([Chap. 4](#)). [Chapters 5](#) and [6](#) are intended to attract especially the researchers focused on metalearning, interested in meta-level analysis of learning algorithms and creating new meta-level algorithms.



<http://www.springer.com/978-3-319-00959-9>

Meta-Learning in Decision Tree Induction

Grąbczewski, K.

2014, XVI, 343 p. 33 illus., Hardcover

ISBN: 978-3-319-00959-9