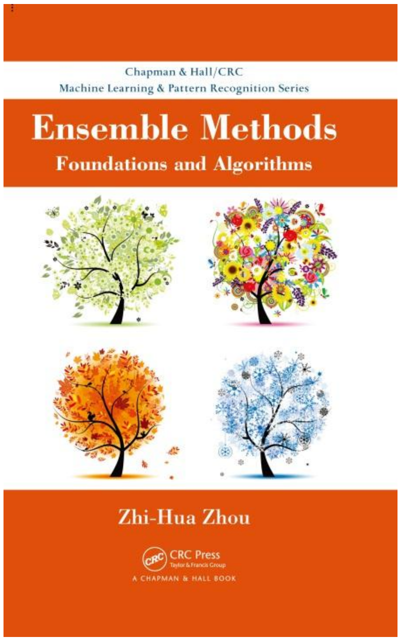
Ensemble Methods: Foundations and Algorithms

BY Zhi-Hua Zhou - ISBN 978-1-439-830031



REVIEWED BY DIRK VAN DEN POEL

nsemble methods train multiple learners and then combine them for use. They have become a hot topic in academia since the 1990s, and are enjoying increased attention in industry. This is mainly based on their generalization ability, which is often much stronger than that of simple/base learners. Ensemble methods are able to boost weak learners, which are even just slightly better than random performance to strong learners, which can make very accurate predictions.

Zhi-Hua Zhou's "Ensemble Methods: Foundations and Algorithms" starts off in Chapter 1 with a brief introduction to the basics, by discussing nomenclature and the basic classifiers including, naive bayes, SVM, k-NN, decision trees, etc.

The real ensemble content kicks off with a discussion of Boosting (Chapter 2), followed by Bagging (Chapter 3). These two chapters form the heart of the book; hence they are discussing the topic in detail. The boosting chapter explains the basic idea, which starts by fitting one learner, and correcting its "mistakes" in subsequent learners. Adaboost is its best known representative of the residual-decreasing methods, which is explained in-depth in Chapter 2. It is an example of a sequential ensemble method. Error bounds of the final combined learner are discussed based on the errors of its weak base learners. Mostly, the book first explains the binary classification problem, and then ventures into multi-class extensions (one-versus-all, one-versus-one approaches), also in this case for multiclass Adaboost. It is well known that the algorithm suffers from noisy data. Hence, the remainder of this chapter mainly focuses on how the algorithm can be made less vulnerable to its weakness to noisy data.

Chapter 3 details the Bagging idea (Boostrap AGGregatING), which is a parallel ensemble method, and lends itself ideally to the possibility of parallel computing. Bagging uses bootstrap sampling (i.e., composing a new dataset of the same size by sampling with replacement from the base dataset). It builds on the idea that the combination of independent base learners will lead to a substantial decrease of errors and therefore, we want to obtain base learners as independent as possible. The bootstrapping leads to a nice side-benefit: Thanks to sampling with replacement, about 37% of the base dataset remains unused, i.e., out-of-bag validation performance can be computed to assess the quality of the learner. Talking about bagging would not be complete without talking about Random Forest, Breiman's random tree ensemble. They can also be found in the book.

Chapter 4 talks about combination methods, which form the basis to achieve strong generalization ability. The author starts with the most prominent form of combination methods: Averaging (simple, weighted, etc.) for regression, and voting (majority, weighted, plurality, etc.) for classification. Next, Stacking (also known as constructing a meta-learner); the idea of stacking is to train the first-level learners using the original training dataset, and then generate a new dataset for training the second-level (meta) learner, where the outputs of the first-level learners are regarded as input features. Next, the author goes on to discuss a number of other combination methods: algebraic methods, Behavior Knowledge Space (BKS) method and decision template method.

Diversity is the foundation on which the performance of ensembles is built. Hence, the book devotes an entire chapter (5) to this topic, providing a lot of information of diversity measures.

Chapter 6 is devoted to ensemble pruning: Instead of using all learners, why not use a subset of them. Generally, it is better to retain some accurate learners together with some not-that-good but complementary learners. The author discusses ordering-based pruning, clustering-based pruning, and optimization-based pruning.

In Chapter 7 the book discusses Clustering Ensembles. These are desired to improve clustering quality, clustering robustness, etc., although their original motivation was to enable knowledge reuse and distributed computing. The author discusses similarity-based methods, graph-based methods, relabeling-based methods, and transformation-based methods.

Finally, Chapter 8 discusses advanced topics such as semi-supervised learning with ensembles, active learning, and class-imbalance learning. In real-world applications, in addition to attaining good accuracy, the comprehensibility of the learned model is also important, because an ensemble aggregates multiple models. Among my favorite parts of the book: A discussion of the alternative ways to achieve this objective: e.g. reduction of the ensemble to a single model.

It is always exciting to read a new book of a prominent researcher in the field. Zhi-Hua Zhou's book certainly qualifies in this category. Discussion in the book starts from a theoretical foundation, but the author also includes many references to successful applications, which makes it a good book both for the researcher and the practitioner. Moreover, this book is not written from a single point of view, but rather includes the view from pattern recognition, data mining as well as (to a lesser extent) statistics.

Important algorithms/approaches are discussed in pseudo-codes, which facilitates the understanding. The author does not just provide the math, but also a clear explanation of the reasoning behind it. The discussion starts with the basic algorithm, and then introduces a number of improvements that have been published in leading scientific journals. At the end of each chapter, there is always a "further readings" section providing hints for literature reading.

What I missed in this book? Some of the statistical methods (logistic regression), references to software and hybrid ensembles. This should be seen as suggestions for a second edition of the book, rather than as real problems. A book is always a compromise. Unlike a website, a book has to be balanced, which means one cannot provide asymmetric depth in the different topics.

In sum, this book deserves a special place in my library. It is well-written, and provides a very clear explanation of the different ensemble approaches including the intuition behind the algorithms why some of them work so well, and most of all, it provides an comprehensive overview of the alternative approaches (as opposed to the academic papers, where it lies scattered in thousands of (small) contributions).

THE BOOK:

ZHI-HUA ZHOU (2012), ENSEMBLE METHODS: FOUNDATIONS AND ALGORITHMS, 236 P. BOCA RATON, FL: CHAPMAN & HALL/CRC.
ISBN: 978-1-439-830031

ABOUT THE REVIEWER:

DIRK VAN DEN POEL
Marketing Analytics at Faculty of Economics and Business Administration, Ghent University, Belgium. Contact him at: dirk.vandenpoel@ugent.be