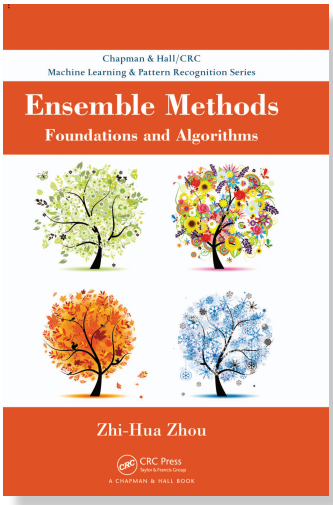
Friedhelm Schwenker
Ulm University, GERMANY

Book Review

Ensemble Methods: Foundations and Algorithms, by Zhi-Hua Zhou (New York: Chapman & Hall/CRC, 2012, 222 pp.) ISBN: 978-1-4398-3003-1.

Learning a model from a finite set of training data is an ill-posed problem. The idea of ensemble learning is to construct a pool of learners and combine them in a smart way into an overall system, rather than to construct a monolithic system. Basically two questions appear in this context. The first question deals with the problem of generating a pool of accurate and diverse base learners, and the second question in this context is, how to combine the outputs of the base learners to achieve a maximal accurate overall system. Ensemble learning has become a popular machine learning approach during the last 20 years, and the series of workshops on Multiple Classifier System founded by Joseph Kittler and Fabio Roli in the year 2000 is one of the most popular forums in the field of ensemble learning.

Professor Zhou's monograph is divided into eight chapters: Chapter 1 of the book issues basic knowledge of machine learning and pattern recognition which is relevant for the rest of the book. The major purpose of this chapter is to explain the terminology used throughout the book, and to introduce the basic concepts of machine learning algorithms. Some of the state-of-the-art learning algorithms are briefly discussed, these are statistical methods for instance linear discriminant analysis, decision trees, Naïve Bayes, K-nearest Neighours, as well as artificial neural networks, support vector machines and kernel-based classifiers. Besides these descriptions of basic learning algorithms, a brief discussion on statistical evaluation tools is given, particularly cross-validation testing and statistical tests to compare results achieved from different algorithms, e.g. paired-t-test, McNemar-test, and Friedman-test are introduced here. Because the terminology of ensemble learning is not fixed in detail yet, the relevant terminology is defined in this chapter. A review of a variety of ensemble learning applications is given to motivate the ensemble approach. This first chapter ends—as any other chapter of this book—with a brief collection of "Further Readings" pointing to papers, monographs, and textbooks relevant to material of the chapter's topics.



Chapter 2 introduces the famous boosting algorithms. Starting point of this chapter is the basic idea of Schapire's algorithm boosting weak learners to strong learners. Schapire's result positively answers the question whether the complexity classes of weak learnable and strong learnable problems are equal. This general boosting procedure is presented in form of a detailed pseudo-code table—almost all algorithms are presented. The main part of this chapter is devoted to the famous Ada-Boost algorithm developed by Freund and Schapire in the mid 1990s. Ada-Boost is perhaps one of the most important approaches in machine learning. At first, Ada-Boost is presented for two-class classification, and then in chapter 2.5, its extension to multi-class classification problems is introduced too. In chapter 2.2, the basic processing steps of Ada-Boost are derived and analyzed. In combination with an illustrative example of the 2D-XOR-problem, this part is very useful to gain a deeper understanding of the Ada-Boost algorithm. Besides these "educational issues" theoretical aspects of Ada-Boost are reviewed in chapter 2.4, in particular the connection between large margin classifiers and Ada-boost as well as the idea to consider Ada-Boost as an interactive optimization process fitting additive models based on the surrogate loss function. Ada-Boost works by learning a set

Digital Object Identifier 10.1109/MCI.2012.2228600
Date of publication: 14 January 2013

It is well known that an ensemble is usually significantly more accurate then a single learner.

of classifiers sequentially where the later classifiers focus on the mistakes of the earlier classifiers, thus if a training example has been incorrectly labelled, the distribution weight increases for this instance and the Ada-Boost algorithm tries to produce this given, but incorrect label. This might degrade the overall performance of the classifier. Originally, Ada-Boost was designed for noise free data, but of course, several practical extensions of Ada-Boost taking into account noisy data have been introduced in the literature. The most important "noise tolerant" implementation of Ada-Boost, namely FilterBoost, MadaBoost, Boosting-by-majority, BrownBoost, and RobustBoost, which have been developed over the last 15 years are included in this chapter.

Chapter 3 focuses on Bagging algorithms. Bagging and Boosting are the most important ensemble learning algorithms. The first part of this chapter introduces the sequential and parallel ensemble creation paradigms, and discusses advantages and disadvantages of these concepts. Whereas Boosting follows the idea of sequential ensemble learning, Bagging belongs to the family of parallel ensemble learning techniques. As for the Ada-Boost algorithm in chapter 2, an illustrative example is given and the behavior of bagged decision trees is discussed, numerous figures and diagrams are presented in order to illustrate the achieved results. Bagging is widely applied in combination with decision tree induction, and a variety of algorithms have been described in the literature. In chapter 3.5, Breiman's Random Forest approach, which is perhaps the most important bagging algorithm, is discussed in detail. Furthermore, random tree ensembles for the density estimation problem are reviewed here in the context of bagged learners.

Chapter 4 "Combination Methods" presents a very general introduction into the field of ensemble learning, and this chapter might be considered as the main part of this book. In the introduction of this chapter the three fundamental benefits of combining multiple learners introduced by Dietterich in the year 2000 at the first Multiple Classifier Systems workshop are reviewed. Subsequently, classifier combination by averaging (chapter 4.2) and voting (chapter 4.3) is presented. These schemes are probably the most important and widely applied combination/aggregation rules for combining classifiers in practice. Furthermore some variants of these two approaches, such as weighted averaging, weighted voting and soft voting are presented and analyzed. Besides these practical aspects of ensembles, the most relevant theoretical results which have been derived for ensemble learning algorithms, for instance the decision boundary analysis introduced by Tumer and Ghosh (1996) and the more recent analysis of the majority voting rule introduced by Narasimhamurthy (2003 and 2005), are presented and discussed in detail. Whereas averaging and voting are fixed aggregation rules, the very general concept of learning the aggregation mapping is another option for the combination of multiple learners, this topic is discussed in chapter 4.4. The Stacking algorithm which was invented by Wolpert (1992) is described, followed by a discussion of the infinite ensemble approach. Subsequently, algebraic methods, including the maximum, minimum and median combination rules, behavior knowledge space and decision template algorithms are discussed. Whereas the algebraic methods can be considered as static fusion schemes the latter two are learnable combiners. Chapter 4.6 is devoted to methods that are utilizing multi learners but cannot be considered to be ensemble algorithms in a strict sense, e.g. error-correcting-output-codes first introduced in Dieterrich and Bakiri (1995), behavior knowledge space, and the mixture of experts invented by Jacobs and colleagues in the early 1990s.

Chapter 5 is dealing with ensemble diversity. Obviously, individual learners of an ensemble should be accurate and different in some sense. Intuitively speaking, learners, for instance classifiers should make their errors on different data sets. The difference or more technical the diversity among the ensemble members is a fundamental issue of ensembles and is discussed in this chapter. First, it is reviewed that the generalization error of an ensemble of regression functions depends on a term that is strongly related to the diversity of the ensemble members. For this, two error decomposition techniques are presented, the error-ambiguity decomposition introduced by Krogh and Vedelsby (1995), and the famous bias-variance-covariance decomposition, introduced as the bias-variance dilemma in the machine learning literature by Geman et al in 1992. In chapter 5.3, various types of diversity measures are defined, classified in pairwise and non-pairwise criteria, and furthermore limitations of these diversity measures are discussed. Although recent developments in information theoretic diversity seems to offer a new direction of research in ensemble diversity, it is highlighted that the right formulation and criterion to measure diversity has not been found so far. The information theoretic diversity approach has been introduced in the multiple classifier systems literature by Brown (2009) and Zhou and Li (2010). Chapter 5.4, reviews these links between information theory and ensemble methods, it starts with a basic introduction to mutual information which is based on the fundamental concept of entropy in classical information theory. Then two different generalizations of mutual information are given. In its classical definition, mutual information is a bivariate measure of two random variables, multivariate

extensions are the interaction information introduced by Mc Gill (1954) and the concept of multi-information diversity introduced by Watanabe (1960). Besides these more theoretical issues of finding the right diversity formulation or definition, heuristic ways to generate diversity in practical applications are reviewed as well. For instance, data sample manipulation or input feature manipulation, manipulation of learning parameters or variation the output representation are described.

Chapter 6 is devoted to the problem of ensemble pruning. Rather than combing the whole set of pre-trained individual learners, the basic idea of ensemble pruning is to select a subset of learners to comprise the final ensemble. The chapter starts with the fundamental proof that ensemble subsets can be better than complete ensembles. The analysis presented here has been shown in the famous paper "Ensembling neural networks: Many could be better than one" authored by Zhou et al in 2002. Subsequently, some practical algorithms for ensemble pruning are reviewed. These algorithms are divided into three categories, ordering-based pruning where the individual learners are ordered according to a pre-specified measure, clustering-based pruning where a number of prototype learners are identified through some type of clustering process, and optimization-based pruning where the ensemble pruning procedure is formulated as an optimization problem with the goal to find a subset of learners that maximizes/minimizes a given objective function. For all three types the recent literature is reviewed in chapters 6.4, 6.5, and 6.6, respectively. In the final chapter 6.7, some advanced pruning strategies are discussed, for instance, the instance-based pruning method where individual learners for the ensemble are selected for each instance separately. Finally the close relation of ensemble pruning to feature selection is highlighted and it is stressed that the famous Viola-Jones face detection algorithms can be interpreted as a pruning method for a Haar-feature based decision stump ensemble.

Ensemble methods have already achieved great success in many real-world tasks.

Supervised learning such as classification and regression is the major scenario for the application of ensemble methods. Chapter 7 is devoted to ensembles of (unsupervised) clustering algorithms. At first, a general introduction to the field of cluster analysis is given, starting with a review of the most relevant algorithm and evaluation criteria for clustering. As for classifier ensembles, the major motivation for ensembling clusterings is to improve the quality and robustness of the clustering results. Besides this, cluster ensembles might be useful in application domains where the data must be processed at distributed location. Similar to the classifier ensembles, the clustering ensemble process is performed in two steps, at first a set of clusterings is generated, with the result that each instance of the data set is labelled by the cluster memberships of each clustering. In a second step, an aggregation function is applied to the given base clusterings in order to achieve the overall clustering. In comparison to classification—where the output structure is defined through the given set of classes /categories—in clustering there is not a natural order of the clustering structure, and therefore the problem of cluster aggregation is much more complex. The main part of the chapter focuses on aggregation rules. For this, these methods are divided into four categories, similarity-based methods, graph methods, relabeling methods, and the so-called transformation-based methods. The relevant state-of-the-art methods and results are reviewed here.

In chapter 8, the book concluded with a review of advanced topics, here recent directions of machine learning research are discussed in the context of ensemble learning. The chapter starts with an introduction to ensembles for semi-supervised and active learning. These two learning approaches can be viewed as partially supervised learning schemes where the learner does not have access to the class labels of the full data set. Cost-sensitive learning , taking into account the total costs, including instance test cost, teacher/expert cost, intervention cost, etc. and class-imbalance learning are introduced and discussed in connection with ensembles in chapter 8.3 and chapter 8.4, respectively. The comprehensibility problem in ensemble learning is mentioned and a few approaches, such as rule extraction and visualization of ensembles, are discussed. As future directions of ensemble learning Professor Zhou highlighted two directions of research—the understanding of ensembles and ensembles in the internet world.

In summary, the present monograph authored by Professor Zhi-Hua Zhou is a valuable contribution to theoretical and practical ensemble learning. The material is very well presented, preliminaries and basic knowledge are discussed in detail, many illustrations and pseudo-code tables help to understand the facts of this interesting field of research. Therefore, the book will become a helpful tool for practitioners working in the field of machine learning or pattern recognition as well as for students of engineering or computer sciences at the graduate and postgraduate level.

I heartily recommend this new book!