Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions

Synthesis Lectures on Data Mining and Knowledge Discovery

Editor
RobertGrossman, *University of Illinois, Chicago*

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions Giovanni Seni and John F. Elder

ISBN: 978-3-031-00771-2 paperback
ISBN: 978-3-031-01899-2 ebook

DOI 10.1007/978-3-031-01899-2

A Publication in the Springer series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #2
Series Editor: Robert Grossman, *University of Illinois, Chicago* Series ISSN
Synthesis Lectures on Data Mining and Knowledge Discovery Print 2151-0067 Electronic 2151-0075

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions

Giovanni Seni
Elder Research, Inc. and Santa Clara University

John F. Elder
Elder Research, Inc. and University of Virginia

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #2

ABSTRACT

Ensemble methods have been called the most influential development in Data Mining and Machine Learning in the past decade. They combine multiple models into one usually more accurate than the best of its components. Ensembles can provide a critical boost to industrial challenges – from investment timing to drug discovery, and fraud detection to recommendation systems – where predictive accuracy is more vital than model interpretability.

Ensembles are useful with all modeling algorithms, but this book focuses on decision trees to explain them most clearly. After describing trees and their strengths and weaknesses, the authors provide an overview of regularization – today understood to be a key reason for the superior performance of modern ensembling algorithms. The book continues with a clear description of two recent developments: *Importance Sampling* (IS) and *Rule Ensembles* (RE). IS reveals classic ensemble methods – bagging, random forests, and boosting – to be special cases of a single algorithm, thereby showing how to improve their accuracy and speed. REs are linear rule models derived from decision tree ensembles. They are the most interpretable version of ensembles, which is essential to applications such as credit scoring and fault diagnosis. Lastly, the authors explain the paradox of how ensembles achieve greater accuracy on new data despite their (apparently much greater) complexity.

This book is aimed at novice and advanced analytic researchers and practitioners – especially in Engineering, Statistics, and Computer Science. Those with little exposure to ensembles will learn why and how to employ this breakthrough method, and advanced practitioners will gain insight into building even more powerful models. Throughout, snippets of code in R are provided to illustrate

the algorithms described and to encourage the reader to try the techniques[1].

The authors are industry experts in data mining and machine learning who are also adjunct professors and popular speakers. Although early pioneers in discovering and using ensembles, they here distill and clarify the recent groundbreaking work of leading academics (such as Jerome Friedman) to bring the benefits of ensembles to practitioners.

The authors would appreciate hearing of errors in or suggested improvements to this book, and may be emailed at seni@datamininglab.com and elder@datamininglab.com. Errata and updates will be available from www.morganclaypool.com

KEYWORDS

ensemble methods, rule ensembles, importance sampling, boosting, random forest, bagging, regularization, decision trees, data mining, machine learning, pattern recognition, model interpretation, model complexity, generalized degrees of freedom

[1]R is an Open Source Language and environment for data analysis and statistical modeling available through the Comprehensive R Archive Network (CRAN). The R system's library packages offer extensive functionality, and be downloaded form http://cran.r-project.org/ for many computing platforms. The CRAN web site also has pointers to tutorial and comprehensive documentation. A variety of excellent introductory books are also available; we particularly like *Introductory Statistics with R* by Peter Dalgaard and *Modern Applied Statistics with S* by W.N. Venables and B.D. Ripley.

To the loving memory of our fathers, Tito and Fletcher

Contents

Acknowledgments

We would like to thank the many people who contributed to the conception and completion of this project. Giovanni had the privilege of meeting with Jerry Friedman regularly to discuss many of the statistical concepts behind ensembles. Prof. Friedman's influence is deep. Bart Goethels and the organizers of ACM-KDD07 first welcomed our tutorial proposal on the topic. Tin Kam Ho favorably reviewed the book idea, Keith Bettinger offered many helpful suggestions on the manuscript, and Matt Strampe assisted with R code. The staff at Morgan & Claypool – especially executive editor Diane Cerra – were diligent and patient in turning the manuscript into a book. Finally, we would like to thank our families for their love and support.

Giovanni Seni and John F. Elder
January 2010

Foreword by Jaffray Woodriff

John Elder is a well-known expert in the field of statistical prediction. He is also a good friend who has mentored me about many techniques for mining complex data for useful information. I have been quite fortunate to collaborate with John on a variety of projects, and there must be a good reason that ensembles played the primary role each time.

I need to explain how we met, as ensembles are responsible! I spent my four years at the University of Virginia investigating the markets. My plan was to become an investment manager after I graduated. All I needed was a profitable technical style that fit my skills and personality (that's all!). After I graduated in 1991, I followed where the data led me during one particular caffeine-fueled, double all-nighter. In a fit of "crazed trial and error" brainstorming I stumbled upon the winning concept of creating one "super-model" from a large and diverse group of base predictive models.

After ten years of combining models for investment management, I decided to investigate where my ideas fit in the general academic body of work. I had moved back to Charlottesville after a stint as a proprietary trader on Wall Street, and I sought out a local expert in the field.

I found John's firm, Elder Research, on the web and hoped that they'd have the time to talk to a data mining novice. I quickly realized that John was not only a leading expert on statistical learning, but a very accomplished speaker popularizing these methods. Fortunately for me, he was curious to talk about prediction and my ideas. Early on,

he pointed out that my multiple model method for investing described by the statistical prediction term, “ensemble.”

John and I have worked together on interesting projects over the past decade. I teamed with Elder Research to compete in the KDD Cup in 2001. We wrote an extensive proposal for a government grant to fund the creation of ensemble-based research and software. In 2007 we joined up to compete against thousands of other teams on the Netflix Prize - achieving a third-place ranking at one point (thanks partly to simple ensembles). We even pulled a brainstorming all-nighter coding up our user rating model, which brought back fond memories of that initial breakthrough so many years before.

The practical implementations of ensemble methods are enormous. Most current implementations of them are quite primitive and this book will definitely raise the state of the art. Giovanni Seni’s thorough mastery of the cutting-edge research and John Elder’s practical experience have combined to make an extremely readable and useful book.

Looking forward, I can imagine software that allows users to seamlessly build ensembles in the manner, say, that skilled architects use CAD software to create design images. I expect that Giovanni and John will be at the forefront of developments in this area, and, if I am lucky, I will be involved as well.

Jaffray Woodriff
CEO, Quantitative Investment Management
Charlottesville, Virginia
January 2010

[Editor’s note: Mr. Woodriff’s investment firm has experienced consistently positive results, and has grown to

be the largest hedge fund manager in the South-East U.S.]

Foreword by Tin Kam Ho

Fruitful solutions to a challenging task have often been found to come from combining an ensemble of experts. Yet for algorithmic solutions to a complex classification task, the utilities of ensembles were first witnessed only in the late 1980's, when the computing power began to support the exploration and deployment of a rich set of classification methods simultaneously. The next two decades saw more and more such approaches come into the research arena, and the development of several consistently successful strategies for ensemble generation and combination. Today, while a complete explanation of all the elements remains elusive, the ensemble methodology has become an indispensable tool for statistical learning. Every researcher and practitioner involved in predictive classification problems can benefit from a good understanding of what is available in this methodology.

This book by Seni and Elder provides a timely, concise introduction to this topic. After an intuitive, highly accessible sketch of the key concerns in predictive learning, the book takes the readers through a shortcut into the heart of the popular tree-based ensemble creation strategies, and follows that with a compact yet clear presentation of the developments in the frontiers of statistics, where active attempts are being made to explain and exploit the mysteries of ensembles through conventional statistical theory and methods. Throughout the book, the methodology is illustrated with varied real-life examples, and augmented with implementations in R-code for the readers to obtain first-hand experience. For practitioners, this handy reference

opens the door to a good understanding of this rich set of tools that holds high promises for the challenging tasks they face. For researchers and students, it provides a succinct outline of the critically relevant pieces of the vast literature, and serves as an excellent summary for this important topic.

The development of ensemble methods is by no means complete. Among the most interesting open challenges are a more thorough understanding of the mathematical structures, mapping of the detailed conditions of applicability, finding scalable and interpretable implementations, dealing with incomplete or imbalanced training samples, and evolving models to adapt to environmental changes. It will be exciting to see this monograph encourage talented individuals to tackle these problems in the coming decades.

Tin Kam Ho
Bell Labs, Alcatel-Lucent
January 2010