Data Mining Algorithms in C++

Data Patterns and Algorithms for Modern Applications

Timothy Masters

Apress®

Data Mining Algorithms in C++

Timothy Masters
Ithaca, New York, USA

ISBN-13 (pbk): 978-1-4842-3314-6 ISBN-13 (electronic): 978-1-4842-3315-3
https://doi.org/10.1007/978-1-4842-3315-3

Library of Congress Control Number: 2017962127

Cover image by Freepik (www.freepik.com)

Managing Director: Welmoed Spahr
Editorial Director: Todd Green
Acquisitions Editor: Steve Anglin
Development Editor: Matthew Moodie
Technical Reviewers: Massimo Nardone and Michael Thomas
Coordinating Editor: Mark Powers
Copy Editor: Kim Wimpsett

Distributed to the book trade worldwide by Springer Science+Business Media New York, 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax (201) 348-4505, e-mail orders-ny@springer-sbm.com, or visit www.springeronline.com. Apress Media, LLC is a California LLC and the sole member (owner) is Springer Science + Business Media Finance Inc (SSBM Finance Inc). SSBM Finance Inc is a **Delaware** corporation.

For information on translations, please e-mail rights@apress.com, or visit www.apress.com/rights-permissions.

Apress titles may be purchased in bulk for academic, corporate, or promotional use. eBook versions and licenses are also available for most titles. For more information, reference our Print and eBook Bulk Sales web page at www.apress.com/bulk-sales.

Any source code or other supplementary material referenced by the author in this book is available to readers on GitHub via the book's product page, located at www.apress.com/9781484233146. For more detailed information, please visit www.apress.com/source-code.

Printed on acid-free paper

Table of Contents

About the Author

Timothy Masters has a PhD in mathematical statistics with a specialization in numerical computing. He has worked predominantly as an independent consultant for government and industry. His early research involved automated feature detection in high-altitude photographs while he developed applications for flood and drought prediction, detection of hidden missile silos, and identification of threatening military vehicles. Later he worked with medical researchers in the development of computer algorithms for distinguishing between benign and malignant cells in needle biopsies. For the past 20 years he has focused primarily on methods for evaluating automated financial market trading systems. He has authored eight books on practical applications of predictive modeling.

- *Deep Belief Nets in C++ and CUDA C: Volume III: Convolutional Nets* (CreateSpace, 2016)
- *Deep Belief Nets in C++ and CUDA C: Volume II: Autoencoding in the Complex Domain* (CreateSpace, 2015)
- *Deep Belief Nets in C++ and CUDA C: Volume I: Restricted Boltzmann Machines and Supervised Feedforward Networks* (CreateSpace, 2015)
- *Assessing and Improving Prediction and Classification* (CreateSpace, 2013)
- *Neural, Novel, and Hybrid Algorithms for Time Series Prediction* (Wiley, 1995)
- *Advanced Algorithms for Neural Networks* (Wiley, 1995)
- *Signal and Image Processing with Neural Networks* (Wiley, 1994)
- *Practical Neural Network Recipes in C++* (Academic Press, 1993)

About the Technical Reviewers

Massimo Nardone has more than 23 years of experience in security, web/mobile development, cloud computing, and IT architecture. His true IT passions are security and Android.

He currently works as the chief information security officer (CISO) for Cargotec Oyj and is a member of the ISACA Finland Chapter board. Over his long career, he has held many positions including project manager, software engineer, research engineer, chief security architect, information security manager, PCI/SCADA auditor, and senior lead IT security/cloud/SCADA architect. In addition, he has been a visiting lecturer and supervisor for exercises at the Networking Laboratory of the Helsinki University of Technology (Aalto University).

Massimo has a master of science degree in computing science from the University of Salerno in Italy, and he holds four international patents (related to PKI, SIP, SAML, and proxies). Besides working on this book, Massimo has reviewed more than 40 IT books for different publishing companies and is the coauthor of *Pro Android Games* (Apress, 2015).

Michael Thomas has worked in software development for more than 20 years as an individual contributor, team lead, program manager, and vice president of engineering. Michael has more than ten years of experience working with mobile devices. His current focus is in the medical sector, using mobile devices to accelerate information transfer between patients and healthcare providers.

Introduction

Data mining is a broad, deep, and frequently ambiguous field. Authorities don't even agree on a definition for the term. What I will do is tell you how I interpret the term, especially as it applies to this book. But first, some personal history that sets the background for this book…

I've been blessed to work as a consultant in a wide variety of fields, enjoying rare diversity in my work. Early in my career, I developed computer algorithms that examined high-altitude photographs in an attempt to discover useful things. How many bushels of wheat can be expected from Midwestern farm fields this year? Are any of those fields showing signs of disease? How much water is stored in mountain ice packs? Is that anomaly a disguised missile silo? Is it a nuclear test site?

Eventually I moved on to the medical field and then finance: Does this photomicrograph of a tissue slice show signs of malignancy? Do these recent price movements presage a market collapse?

All of these endeavors have something in common: they all require that we find variables that are meaningful in the context of the application. These variables might address specific tasks, such as finding effective predictors for a prediction model. Or the variables might address more general tasks such as unguided exploration, seeking unexpected relationships among variables—relationships that might lead to novel approaches to solving the problem.

That, then, is the motivation for this book. I have taken some of my most-used techniques, those that I have found to be especially valuable in the study of relationships among variables, and documented them with basic theoretical foundations and well-commented C++ source code. Naturally, this collection is far from complete. Maybe Volume 2 will appear someday. But this volume should keep you busy for a while.

You may wonder why I have included a few techniques that are widely available in standard statistical packages, namely, very old techniques such as maximum likelihood factor analysis and varimax rotation. In these cases, I included them because they are useful, and yet reliable source code for these techniques is difficult to obtain. There are times when it's more convenient to have your own versions of old workhorses, integrated

into your own personal or proprietary programs, than to be forced to coexist with canned packages that may not fetch data or present results in the way that you want.

You may want to incorporate the routines in this book into your own data mining tools. And that, in a nutshell, is the purpose of this book. I hope that you incorporate these techniques into your own data mining toolbox and find them as useful as I have in my own work.

There is no sense in my listing here the main topics covered in this text; that's what a table of contents is for. But I would like to point out a few special topics not frequently covered in other sources.

- *Information theory* is a foundation of some of the most important techniques for discovering relationships between variables, yet it is voodoo mathematics to many people. For this reason, I devote the entire first chapter to a systematic exploration of this topic. I do apologize to those who purchased my *Assessing and Improving Prediction and Classification* book as well as this one, because Chapter 1 is a nearly exact copy of a chapter in that book. Nonetheless, this material is critical to understanding much later material in this book, and I felt that it would be unfair to almost force you to purchase that earlier book in order to understand some of the most important topics in this book.
- *Uncertainty reduction* is one of the most useful ways to employ information theory to understand how knowledge of one variable lets us gain measurable insight into the behavior of another variable.
- *Schreiber's information transfer* is a fairly recent development that lets us explore causality, the directional transfer of information from one time series to another.
- *Forward stepwise selection* is a venerable technique for building up a set of predictor variables for a model. But a generalization of this method in which ranked sets of predictor candidates allow testing of large numbers of combinations of variables is orders of magnitude more effective at finding meaningful and exploitable relationships between variables.

- *Simple modifications* to relationship criteria let us detect profoundly nonlinear relationships using otherwise linear techniques.
- Now that extremely fast computers are readily available, *Monte Carlo permutation tests* are practical and broadly applicable methods for performing rigorous statistical relationship tests that until recently were intractable.
- *Combinatorially symmetric cross validation* as a means of detecting overfitting in models is a recently developed technique, which, while computationally intensive, can provide valuable information not available as little as five years ago.
- Automated selection of variables suited for predicting a given target has been routine for decades. But in many applications you have a choice of possible targets, any of which will solve your problem. Embedding target selection in the search algorithm adds a useful dimension to the development process.
- *Feature weighting as regularized energy-based learning* (FREL) is a recently developed method for ranking the predictive efficacy of a collection of candidate variables when you are in the situation of having too few cases to employ traditional algorithms.
- Everyone is familiar with *scatterplots* as a means of visualizing the relationship between pairs of variables. But they can be generalized in ways that highlight relationship anomalies far more clearly than scatterplots. Examining discrepancies between joint and marginal distributions, as well as the contribution to mutual information, in regions of the variable space can show exactly where interesting interactions are happening.
- Researchers, especially in the field of psychology, have been using *factor analysis* for decades to identify hidden dimensions in data. But few developers are aware that a frequently ignored byproduct of maximum likelihood factor analysis can be enormously useful to data miners by revealing which variables are in redundant relationships with other variables and which provide unique information.

- Everyone is familiar with using correlation statistics to measure the degree of relationship between pairs of variables, and perhaps even to extend this to the task of clustering variables that have similar behavior. But it is often the case that variables are strongly contaminated by noise, or perhaps by external factors that are not noise but that are of no interest to us. Hence, it can be useful to cluster variables *within the confines of a particular subspace* of interest, ignoring aspects of the relationships that lie outside this desired subspace.
- It is sometimes the case that a collection of time-series variables are coherent; they are impacted as a group by one or more underlying drivers, and so they change in predictable ways as time passes. Conversely, this set of variables may be mostly independent, changing on their own as time passes, regardless of what the other variables are doing. Detecting when your variables *move from one of these states* to the other allows you, among other things, to develop separate models, each optimized for the particular condition.

I have incorporated most of these techniques into a program, DATAMINE, that is available for free download, along with its user's manual. This program is not terribly elegant, as it is intended as a demonstration of the techniques presented in this book rather than as a full-blown research tool. However, the source code for its core routines that is also available for download should allow you to implement your own versions of these techniques. Please do so, and enjoy!