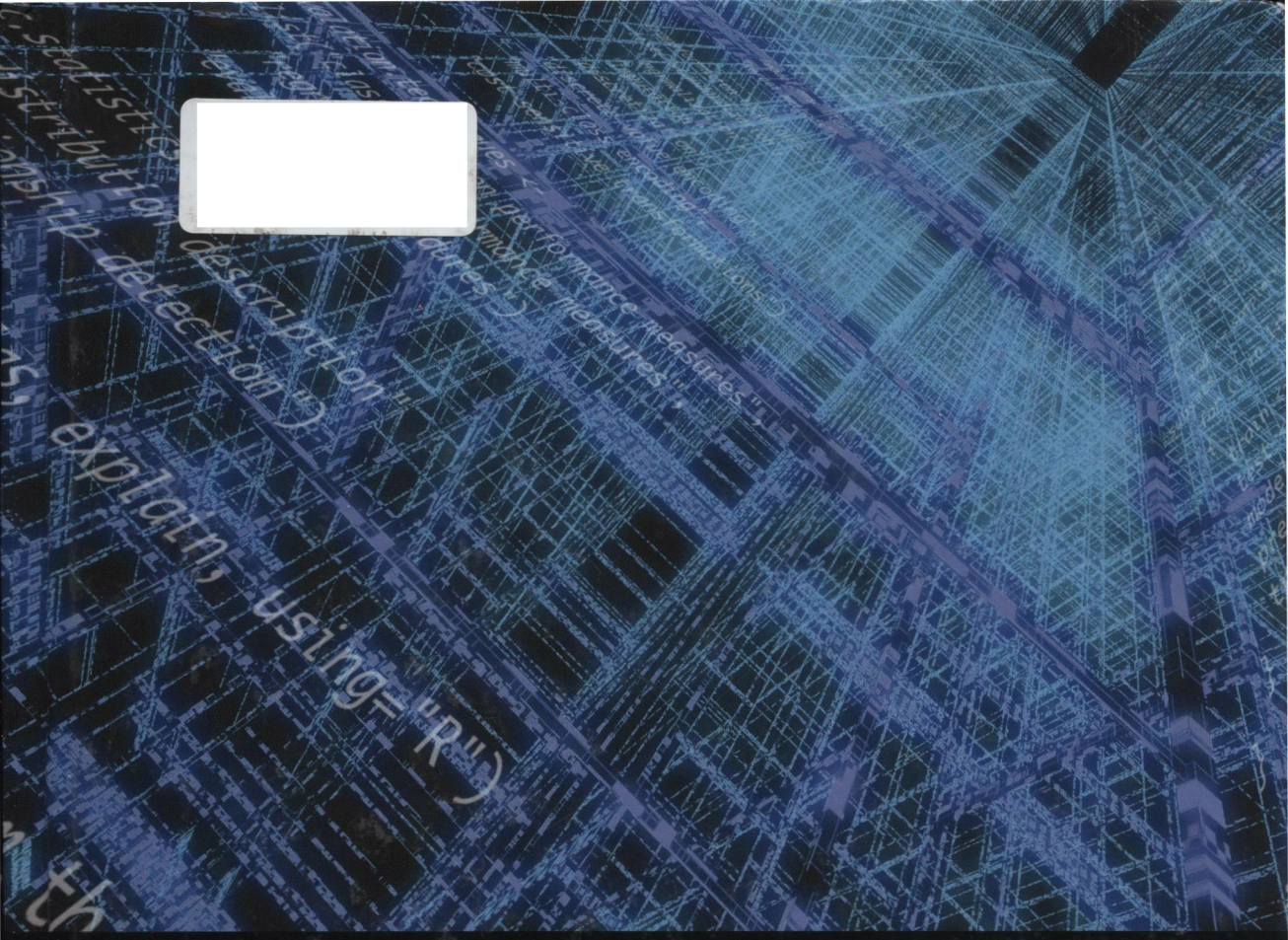

DATA MINING ALGORITHMS

EXPLAINED USING R

Paweł Cichosz

with website

WILEY

Data Mining Algorithms: Explained Using R

Paweł Cichosz

Department of Electronics and Information Technology
Warsaw University of Technology
Poland

WILEY

This edition first published 2015

Registered office: John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Cichosz, Pawel, author.
Data mining algorithms : explained using R / Pawel Cichosz.
pages cm
Summary: "This book narrows down the scope of data mining by adopting a heavily modeling-oriented perspective" – Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-118-33258-0 (hardback)
1. Data mining. 2. Computer algorithms. 3. R (Computer program language) I. Title.
QA76.9.D343C472 2015
006.3′12–dc23

2014036992

A catalogue record for this book is available from the British Library.

ISBN: 9781118332580

Typeset in 10/12pt Times by Laserwords Private Limited, Chennai, India
Printed and bound in Singapore by Markono Print Media Pte Ltd

Data Mining Algorithms

To my wife, Joanna, and my sons,
Grzegorz and Łukasz

Acknowledgements

With the rise and rapidly growing popularity of online idea sharing methods, such as blogs and wikis, traditional books are no longer the only way of making large portions of text available to a wide audience. The former are particularly suitable for collaborative or social writing and readings undertakings, often with mixed reader–writer roles of particular participants. For individual writing and reading efforts the traditional book form (although not necessarily tied to the paper media) still remains the best approach. On one hand, it clearly assigns full and exclusive responsibility for the contents to the author, with no easy excuses for errors and other deficiencies. On the other hand, there are several other people engaged in the publishing process who help to give the book its final shape and protect the audience against a totally flawed work.

As the author of this book, I feel indeed totally responsible for all its imperfections, only some which I am aware of, but I have no doubts that there are many more of them. With that being said, several people from the editorial and production team worked hard to make the imperfect outcome of my work worth publishing. My thanks go, in particular, to Richard Davies, Prachi Sinha Sahay, Debbie Jupe, and Kay Heather from Wiley for their encouragement, support, understanding, and reassuring professionalism at all stages of writing and production. Radhika Sivalingam, Lincy Priya, and Yogesh Kukshal did their best to transform my manuscript into a real book, meeting publication standards. I believe there are others who contributed to this book's production that I am not even aware of and I am grateful to them all, also.

I was thoughtless enough to share my intention to write this book with my colleagues from the Artificial Intelligence Applications Research Group at the Warsaw University of Technology. While their warm reception of this idea and constant words of encouragement were extremely helpful, I wished I had not done that many times. It would have been so much easier to give up if I had kept this in secret. Perhaps the ultimate reason why I continued to work despite hesitations is that I knew they would keep asking and I would be unable to find a good excuse. Several thoughts expressed in this book were shaped by discussions during our group's seminar meetings. Interacting with my colleagues from the analytics teams at Netezza Poland, IBM Poland, and iQor Poland, with which I had an opportunity to work on some data mining projects at different stages of writing the book, was also extremely stimulating, although the contents of the book have no relationships with the projects I was involved with.

I owe special thanks to my wife and two sons, who did not directly contribute to the contents of this book, but made it possible by allowing me to spend much of my time that should be normally devoted to them on this work and providing constant encouragement. If you guys can read these thanks in a published copy of the book, then it means it is all over at last and we will hopefully get back to normal life.

Preface

Data mining

Data mining has been a rapidly growing field of research and practical applications during the last two decades. From a somewhat niche academic area at the intersection of machine learning and statistics it has developed into an established scientific discipline and a highly valued branch of the computing industry. This is reflected by data mining becoming an essential part of computer science education as well as the increasing overall awareness of the term "data mining" among the general (not just computing-related) academic and business audience.

Scope

Various definitions of data mining may be found in the literature. Some of them are broad enough to include all types of data analysis, regardless of the representation and applicability of their results. This book narrows down the scope of data mining by adopting a heavily modeling-oriented perspective. According to this perspective the ultimate goal of data mining is delivering *predictive models*. The latter can be thought of as computationally represented chunks of knowledge about some domain of interest, described by the analyzed data, that are capable of providing answers to queries *transcending the data*, i.e., such that cannot be answered by just extracting and aggregating values from the data. Such knowledge is discovered from data by capturing and generalizing useful relationship patterns that occur therein.

Activities needed for creating predictive models based on data and making sure that they meet the application's requirements fall in the scope of data mining as understood in this book. Analytical activities which do not contribute to model creation – although they may still deliver extremely useful results – remain therefore beyond the scope of our interest. This still leaves a lot of potential contents to be covered, including not only modeling algorithms, but also techniques for evaluating the quality of predictive models, transforming data to make modeling algorithms easier to apply or more likely to succeed, selecting attributes most useful for model creation, and combining multiple models for better predictions.

Modeling view

The modeling view of data mining is by no means unique for this book. It is actually the most natural and probably the most wide-spread view of data mining. Nevertheless, it deserves

some more attention in this introductory discussion, which is supposed to let the reader know what this book is about. In particular, it is essential to underline – and it will be repeatedly underlined on several other occasions throughout the book – that a useful data mining model is not merely a description of some patterns discovered in the data. In other words, it does not only and not mainly represent knowledge about the data, but also – and much more importantly – knowledge about the domain from which the data originates.

The *domain* can be considered a set of entities from the real world about which knowledge is supposed to be delivered by data mining. These can be people (such as customers, employees, patients), machines and devices (such as car engines, computers, or ATMs), events (such as car failures, purchases, or bank transactions), industrial processes (such as manufacturing electronic components, energy production, or natural resources exploitation), business units (such as stores or corporate departments), to name only a few typical possibilities. Such real-world entities – in this book referred to as *instances* – are, usually incompletely and imperfectly, described by a set of features – in this book referred to as *attributes*. A *dataset* is a subset of the domain, described by the set of available attributes, usually – assuming a tabular data representation – with rows corresponding to instances and columns corresponding to attributes. Data mining can then be viewed as an analytic process that uses one or more available datasets from the same domain to create one or more *models* for the domain, i.e., models that can be used to answer queries not just about instances from the data used for model creation, but also about any other instances from the same domain. More directly and technically, speaking, if some attributes are generally available (observable) and some attributes are only available on a limited dataset (hidden), then models can often be viewed as delivering predictions of hidden attributes wherever their true values are unavailable. The unavailable attribute values to be predicted usually represent properties or quantities that are hard and costly to determine, or (more typically) that become known later than are needed. The latter justifies the term "prediction" used when referring to a model's output. The attribute to be predicted is referred to as the *target attribute*, and the observable attributes that can be used for prediction are referred to as the *input attributes*.

Tasks

The most common types of predictive models – or queries they can be used to answer – correspond to the following three major data mining tasks.

Classification. Predicting a discrete target attribute (representing the assignment of instances to a fixed set of possible classes). This could be distinguishing between good and poor customers or products, legitimate and fraudulent credit card transactions or other events, assigning failure types and recommended repair actions to faulty technical devices, etc.

Regression. Predicting a numeric target attribute which represents some quantity of interest. This could be an outcome or a parameter of an industrial process, an amount of money earned or spent, a cost or gain due to a business decision, etc.

Clustering. Predicting the assignment of instances to a set of similarity-based clusters. Clusters are not predetermined, but discovered as part of the modeling process, to achieve possibly high intracluster similarity and possibly low intercluster similarity.

Most real-world data mining projects include one or more instantiations of these three generic tasks. Similarly, most of data mining research contributes, modifies, or evaluates algorithms for these three tasks. These are also the tasks on which this book is focused.

Origin

Data mining techniques have their roots in two fields: machine learning and statistics. With the former traditionally addressing the issue of acquiring knowledge or skill from supplied training information and the latter the issue of describing the data as well as identifying and approximating relationships occurring therein, they both have contributed modeling algorithms. They have also become increasingly closely related, which makes it difficult and actually unnecessary to put hard separating boundaries between them. With that being said, their common terminological and notational conventions remain partially different, and so do background profiles of researchers and practitioners in these fields. Wherever this difference matters, this book is much closer to machine learning than statistics, to the extent that the description of "strictly statistical" techniques – appearing rather sparingly – may be found oversimplified by statisticians. In particular, the formulations of major data mining tasks in Chapter 1 assume the inductive learning perspective.

The brief discussion of the modeling view of data mining presented in the previous section makes it possible to encounter this book's bias toward machine learning for the first time. The terms "domain," "instance," "attribute," and "dataset," in particular, have their counterparts that are more common in statistics, such as "population," "observation," "variable," and "sample."

Motivation

The book is intended to be a practical, technically oriented guide to data mining algorithms, focused on clearly explaining their internal operation and properties as well as major principles of their application. According to the general perspective of data mining adopted by the book, it encompasses all analytic processes performed to produce predictive models from available data and verify whether and to what extent they meet the application's requirements. The book will cover the most important algorithms for building classification, regression, and clustering models, as well as techniques used for attribute selection and transformation, model quality evaluation, and creating model ensembles.

The book will hopefully appeal to the reader, either already familiar with data mining to some extent or just approaching the field, by its practical and technical, utility-driven perspective, making it possible to quickly start gaining his or her own hands-on experience. The reader will be given an opportunity to become familiar with a number of data mining algorithms, presented in a systematic, coherent, and relatively easy to follow way. By studying their description and examples the reader will learn how they work, what properties they exhibit, and how they can be used.

The book is not intended to be a "data mining bible" providing a complete coverage of the area, but rather to selectively focus on a number of algorithms that:

- are known to work well for the most common data mining tasks,
- are good representatives of typical data mining techniques,

- can be well explained to the general technically educated audience without an excessive required mathematical and computing background,
- can be used to illustrate good practices as well as caveats of data mining.

It is not supposed to be a business-oriented manager's guide to data mining or a bird's eye-perspective overview of the field, either. Topics covered by the book are discussed in a technical way, with a level of detail believed to be adequate for most practical needs, albeit not overwhelming. This includes the presentation of the internal mechanisms, properties, and usage scenarios of algorithms that are not extremely complex mathematically or implementationally and offer the potential of excellent results in many applications, but may need expertise and experience to be used fruitfully. The ambition of the book is to help the reader develop that expertise and experience.

The book's technical and practical orientation, with very limited theoretical background, but a relatively high level of detail on algorithm internal operation and application principles, makes it appropriate for a mixed audience consisting of

- students of computer science and related fields,
- researchers working on experimental or applied research projects in any area where data analysis capabilities are used,
- analysts and engineers working with data and creating or using predictive models.

The book should be particularly appealing to computer scientists and programmers due to its extensive use of R code examples, as explained below.

While the little-background assumption makes the book suitable as an introductory text, the level of detail and precision puts it actually on an advanced or semi-advanced level, since on many occasions – whenever it is justified by practical utility – it discusses issues that tend too be overlooked or taken lightly in typical introductions to data mining.

Organization

The book is divided into the following parts:

Part I. Preliminaries.
Part II. Classification.
Part III. Regression.
Part IV. Clustering.
Part V. Getting better models.

Part I contains two chapters, as summarized below.

Chapter 1: Tasks. This chapter introduces the major data mining tasks algorithms for which are presented in the book: classification, regression, and clustering.

Chapter 2: Basic statistics. This chapter is be devoted to simple techniques for performing data exploration tasks, usually referred to as basic statistics, that are often applied before any modeling algorithms or used internally by some modeling algorithms.

Part II contains five chapters listed below.

Chapter 3: Decision trees. This chapter presents algorithms for creating decision tree classification models, shortly referred to as decision tree algorithms.

Chapter 4: Naïve Bayes classifier. This chapter presents arguably the simplest useful classification algorithm, the naïve Bayes classifier.

Chapter 5: Linear classification. This chapter is devoted to classification algorithms that adopt a linear model representation. Since they are largely based on linear regression algorithms, forward references to Chapter 8 are unavoidable.

Chapter 6: Misclassification costs. This chapter systematically discusses the issue of nonuniform misclassification costs in the classification task and techniques that can be used to create cost-sensitive classification models.

Chapter 7: Classification model evaluation. This chapter is devoted to techniques used to evaluate classification models: performance measures serving as model quality indicators on a particular dataset, and evaluation procedures used to reliably estimate their expected values on new data. Since the presented evaluation procedures are applicable to regression and clustering models as well, the extensive discussion presented in this chapter makes it possible to keep the chapters on regression and clustering model evaluation considerably shorter.

The contents of Part III is summarized below.

Chapter 8: Linear regression. This chapter presents regression algorithms that employ a linear model representation and parameter estimation techniques used to find model parameters.

Chapter 9: Regression trees. This chapter presents regression tree algorithms, which are decision trees adapted for the regression task. This makes it possible to refer to Chapter 3 extensively and focus mostly on regression-specific issues.

Chapter 10: Regression model evaluation. This chapter is devoted to techniques used to assess the quality of regression models. It will focus mostly on regression model performance measures, since evaluation procedures applicable to regression models are the same as those for classification models presented in Chapter 7.

Part IV contains four chapters listed below.

Chapter 11: Dissimilarity measures. This chapter presents several dissimilarity and similarity measures used for clustering.

Chapter 12: k-Centers clustering. This chapter is devoted to the popular k-centers family of clustering algorithms.

Chapter 13: Hierarchical clustering. This chapter presents algorithms for creating hierarchical clustering models.

Chapter 14: Clustering model evaluation. This chapter discusses several quality measures used for clustering model evaluation.

Part V is more heterogenic than the preceding parts, as it covers a set of diverse techniques that can be used to improve the quality of classification, regression, or clustering models. The corresponding list of chapters is presented below.

Chapter 15: Model ensembles. This chapter reviews ensemble modeling algorithms that combine multiple models for the same task classification or regression task for better predictive power.

Chapter 16: Kernel methods. This chapter discusses the possibility of improving the capabilities of liner classification and regression models by employing kernel functions to perform implicit nonlinear data transformation into a higher-dimensional space. This is also an opportunity to discuss the support vector machines and support vector regression algorithms, with which kernel functions are most often combined.

Chapter 17: Attribute transformation. This chapter presents selected techniques used for transforming data prior to applying modeling algorithms, to make them easier to apply and more likely to deliver good models.

Chapter 18: Discretization. This chapter addresses one specific type of data transformation, the discretization of continuous attributes, that is particularly often applied and for which several different algorithms have been developed.

Chapter 19: Attribute selection. This chapter is devoted to attribute selection algorithms, used to select a subset of available attributes with the highest predictive utility.

The last chapter in the book, Chapter 20, is placed outside the division into parts, since it contains data mining case studies that put the algorithms covered in the previous chapters at real work. Existing R implementations and publicly available datasets will be used to demonstrate the process of searching for the possibly best model step by step. This will include data transformation whenever necessary or useful, attribute selection, modeling algorithm application with parameter tuning, and model evaluation.

Notation

Different families of data mining algorithms are often presented in the literature using different notational conventions. Even for different descriptions of the same or closely related algorithms, it is not uncommon to adopt different notation. This is hardly acceptable in a single book, though, even if it covers a variety of techniques, sometimes with completely different origin. To keep the notation consistent throughout the book, it sometimes departs considerably from notational standards typical for particular algorithms and subareas of data mining. The adopted unified notation may therefore sometimes appear nonstandard or even awkward to readers accustomed to other conventions. It should be still much less confusing than changing notation from chapter to chapter.

This book's notational conventions are governed by the modeling view of data mining presented above and similarly biased toward machine learning rather than statistics. They include, in particular, explicit references to instances as elements of the domain, attributes as functions assigning values to instances, and datasets as subsets of the domain. For datasets there are also subscripting conventions used to refer to their subsets satisfying equality or inequality conditions for a particular attribute. Most of them are introduced early in the book, in Chapters 1 and 2, giving the reader enough time to get used to them before passing to modeling algorithms. They are all collected and explained in Appendix A for easy reference.

R code examples

One of primary features of this book is the extensive use of examples, which is necessary to make the desired combination of depth, precision, and readability possible. All of these examples contain R code snippets, in most cases sufficiently simple to be at least roughly comprehensible without prior knowledge of the language. There are two types of these examples.

Algorithm operation illustrations. These are numbered examples included in book sections devoted to particular algorithms or single major steps of more complex algorithms, supposed to help explain the details of internal algorithm calculations. Most of them present R implementations of either complete algorithms (for simple ones) or single steps thereof (for more complex ones). This supplements the natural language, pseudocode, or math formula algorithm description, adding some more clarification, specificity, and appeal. While serving the illustrative purpose mostly, usually inefficient and suited to single simple usage scenarios only, some portions of this example code may actually be useful for practical applications as well. This is at least likely for functions which have no direct counterparts in existing R packages. The insufficient efficiency and flexibility of these illustrative implementations will hopefully encourage the readers to develop more useful modified versions thereof.

Case studies. These bigger examples grouped in the book's final chapter are more realistic demonstrations of performing data mining tasks using standard R implementations of the algorithms described in the book and publicly available datasets. Their purpose is not to explain how particular algorithms work – which is the responsibility of the other chapters and their examples – or how particular functions should be called – which is easy to learn from their R help pages. Instead, their primary focus is on how to solve a given task using one or more available algorithms. They present the process of searching for good quality predictive models that may include data transformation, attribute selection, model building, and model evaluation. They will hopefully encourage the readers to approach similar tasks by their own.

Of those, the first category occupies much more space and corresponds to about 10 times more code lines. It can also be considered a distinctive feature of this book. This is because, while it is a relatively common practice for contemporary books on data mining or statistics to use R (or another analytic toolbox) to provide algorithm usage demonstrations or case studies, the idea of R code snippets for algorithm operation illustrations is probably relatively uncommon.

R is an increasingly popular programming language and environment for data analysis, sometimes referred to as the "lingua franca" of this domain, with a huge set of contributed packages available from the CRAN repository,[1] providing implementations of various analytic algorithms and utility functions. The book uses the R language extensively as a pedagogical tool, but does not teach it nor requires the readers to learn it. This is because the example code can be run and the results can looked up with barely any knowledge of R. Elementary knowledge of any general-purpose programming language augmented by a small set of R-specific

[1] *Comprehensive R Archive Network* (http://cran.r-project.org/web/packages).

features, such as its ubiquitous vectorization and logical indexing, should be sufficient to get at least some rough understanding of most of the presented example code snippets.

However, investing some effort into learning the basics of R will definitely pay off, making it possible to fully exploit the illustrative value of this book's examples. They will hopefully encourage at least some readers to study readily available tutorials and provide useful starting points for such self-study. Making the reader familiar with R is not therefore the purpose of the book, but may become its beneficial side effect.

Needless to say, R code algorithm illustrations and case studies need data. In some examples tiny and totally unrealistic datasets are used to make it possible to manually verify the results. In some other examples slightly larger artificial datasets are generated. On several occasions, however, publicly available real datasets are used, available in CRAN packages and originating from the *UCI Machine Learning Repository*.[2] These are listed in Appendix C.

It is not uncommon for some R functions defined in examples to be reused by examples in other chapters. This is due to the natural relationships among data mining algorithms, some having common operations and some being typically used in combinations. To make it easier to run the example code with such dependences, all potentially reusable functions defined in particular chapters are grouped into corresponding R packages. They all share the same `dmr` name prefix and are available from the book's website.

These packages, referred to as DMR packages thereafter, should be thought of as simple containers for example functions and in many respects they do not meet commonly adopted R package standards. The documentation is particularly lacking, limited to references to the books section and example numbers, but this is forgivable given the fact that they are not distributed as standalone software tools, but as a supplementary material for the book. Some frequently reused utility functions that have no illustrative value on their own and therefore are not included in any of this book's examples, are grouped in the `dmr.util` package. In the first example of each chapter all packages used in subsequent examples – both DMR and CRAN ones – are explicitly loaded. Additionally, whenever a function from another chapter is mentioned, the corresponding example and package containing its definition are indicated in a margin note, as demonstrated here for the `err` function. Appendix B contains the list of all DMR and CRAN packages used in this book.

EX. 7.2.1
`dmr.claseval`

Since the primary role of the code is didactic and illustrative, it is written without any care for efficiency and error handling, and it may not always demonstrate a good R programming style. Testing was mostly limited to the presented example calls. Since the book's chapters were created over an extended period of time, there are some noticeable inconsistencies between their R code illustrations. With all those fully justified disclaimers, the R code snippets are believed to deserve the space they occupy in the book and – while some readers may choose to skip them while reading – the definitely recommended way to use the book is to stop at each example not only to run the code and inspect the results, but ideally also to understand how it matches the preceding text or equations.

[2] http://archive.ics.uci.edu/ml