Cluster and Classification Techniques for the Biosciences

Recent advances in experimental methods have resulted in the generation of enormous volumes of data across the life sciences. Hence clustering and classification techniques that were once predominantly the domain of ecologists are now being used more widely. This book provides an overview of these important data analysis methods, from long-established statistical methods to more recent machine learning techniques. It aims to provide a framework that will enable the reader to recognise the assumptions and constraints that are implicit in all such techniques. Important generic issues are discussed first and then the major families of algorithms are described. Throughout the focus is on explanation and understanding and readers are directed to other resources that provide additional mathematical rigour when it is required. Examples taken from across the whole of biology, including bioinformatics, are provided throughout the book to illustrate the key concepts and each technique's potential.

ALAN H. FIELDING is Senior Lecturer in the Division of Biology at Manchester Metropolitan University.

Cluster and Classification Techniques for the Biosciences

ALAN H. FIELDING

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, UK

Published in the United States of America by Cambridge University Press,
New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521852814

First published 2007

Printed in the United Kingdom at the University Press, Cambridge

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

Library of Congress Cataloging in Publication Data

Fielding, Alan,
Cluster and classification techniques for the biosciences / Alan H. Fielding.
p. ; cm.
Includes bibliographical references and index.
ISBN-13: 978-0-521-85281-4 (hardback)
ISBN-10: 0-521-85281-1 (hardback)
ISBN-10: 0-521-61800-2 (pbk.)
1. Biology–Data processing. 2. Biology–Classification. 3. Cluster analysis. I. Title.
[DNLM: 1. Cluster Analysis. 2. Biometry–methods. 3. Classification–methods. 4. Data Interpretation, Statistical. 5. Multivariate Analysis. WA 950 F459c 2007]
QH324.2F537 2007
570.285–dc22 2006027983

ISBN-13 978-0-521-85281-4 hardback
ISBN-10 0-521-85281-1 hardback

ISBN-13 978-0-521-61800-7 paperback
ISBN-10 0-521-61800-2 paperback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party Internet websites referred to in this book, and does not guarantee that any content on such Web sites, or will remain, accurate or appropriate.

Dedication

This book is dedicated to the memory of Derek Ratcliffe. Derek, who was one of Britain's most important biologists, died during the writing of this book. I had the great pleasure of helping him in some very small ways with data analyses and he kindly provided a reference for my Leverhulme Fellowship award. His humility, breadth, insight and commitment are a great inspiration.

Contents

Cambridge University Press
978-0-521-61800-7 - Cluster and Classification Techniques for the Biosciences
Alan H. Fielding
Frontmatter
More information

Cambridge University Press
978-0-521-61800-7 - Cluster and Classification Techniques for the Biosciences
Alan H. Fielding
Frontmatter
More information

Preface

When I was originally asked to write this book I said no, several times. My reticence was a consequence of workload and not because I thought there was no need for the book. However, Katrina Halliday from Cambridge University Press persisted and eventually I gave in. However, I can only blame myself for any errors or omissions in the text.

Katrina asked me because of my editorship of an earlier machine learning book and favourable comments about some of my web material, in particular a postgraduate multivariate statistics unit. My interest in multivariate statistics arose from my research as a conservation biologist. As part of this research I spent time trying to develop predictive species distribution models, which led to my exploration of two additional topics: machine learning methods as alternatives to statistical approaches; and how to measure the accuracy of a model's predictions.

If you are not an ecologist you may be thinking that there will be little of value for you in the book. Hopefully, the contents will alleviate these fears. My multivariate statistics unit was delivered to a diverse group of students including biomedical scientists, so I am used to looking beyond ecology and conservation. In my experience there is much to be gained by straying outside of the normal boundaries of our research. Indeed my own research into the accuracy of ecological models drew greatly on ideas from biomedical research. At a fundamental level there is a great deal of similarity between a table of species abundance across a range of sites and a table of gene expression profiles across a range of samples. The subject-specific information is obviously important in devising the questions to be answered and the interpretation of results, but it is important to be aware that someone in another discipline may already have devised a suitable approach for a similar problem. Classifiers are in use in many biological and related disciplines; unfortunately there seems to be little cross-fertilisation. Throughout the book I have attempted to draw examples from a wide range of sources including ecology and bioinformatics. I hope that readers will recognise that other disciplines may already have a solution to their problem, or at least a catalogue of the difficulties and pitfalls!

As you read this text you may notice a low equation and symbol count. While this opens the text up to some obvious criticism I feel that the additional accessibility benefits outweigh the deficiencies. I am ready, if somewhat nervous, to face the reviewers' criticisms. While there are many good printed and online resources, which provide the necessary theoretical detail, I think that there is a shortage of overview resources which attempt to provide a framework that is not clouded by too much technical detail. However, the readers are expected to have some basic understanding of statistical methods. What many biologists need, and what I hope to achieve, is support in deciding if their problem can be investigated using a clustering or classification algorithm. I hope to do this by providing general guidelines and examples drawn from across biology. Given the ephemeral nature of many web pages there are relatively few web links in the text. The main exceptions are classic 'papers' and data or software sources.

During the writing several colleagues have commented on parts of the text. I am particularly grateful to Paul Craze, Les May and Emma Shaw. My research colleagues (Paul Haworth, Phil Whitfield and David McLeod) kept me entertained and revitalised during the several research meetings that we had while I was writing this book. However, I fear they may be responsible for physiological damage!

Finally, this book could not have been written without the continuing support of my wife Sue and daughter Rosie.