



R and DATA MINING

Examples and Case Studies

Yanchang Zhao

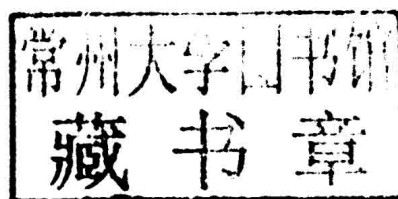


R and Data Mining

Examples and Case Studies

Yanchang Zhao

RDataMining.com



AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK
OXFORD • PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE
SYDNEY • TOKYO

Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA
225 Wyman Street, Waltham, MA 02451, USA
32 Jamestown Road, London NW17BY, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2013

© 2013 Yanchang Zhao. Published by Elsevier Inc. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher.

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting *Obtaining permission to use Elsevier material*.

Notice

No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made.

Library of Congress Cataloging-in-Publication Data

Application submitted

British Library Cataloguing-in-Publication Data

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

ISBN: 978-0-123-96963-7

For information on all Academic Press publications visit
our website at store.elsevier.com

Printed and bound in USA

13 14 15 16 10 9 8 7 6 5 4 3 2 1

Working together to grow
libraries in developing countries

www.elsevier.com | www.bookaid.org | www.sabre.org

ELSEVIER BOOK AID International Sabre Foundation

R and Data Mining

To Yanbo, Michael and Lucas for your love and encouragement

List of Abbreviations

ARIMA	Autoregressive integrated moving average
ARMA	Autoregressive moving average
AVF	Attribute value frequency
CLARA	Clustering for large applications
CRISP-DM	Cross industry standard process for data mining
DBSCAN	Density-based spatial clustering of applications with noise
DTW	Dynamic time warping
DWT	Discrete wavelet transform
GLM	Generalized linear model
IQR	Interquartile range, i.e., the range between the first and third quartiles
LOF	Local outlier factor
PAM	Partitioning around medoids
PCA	Principal component analysis
STL	Seasonal-trend decomposition based on Loess
TF-IDF	Term frequency-inverse document frequency

Contents

List of Figures	xi
List of Abbreviations	xv
1 Introduction	1
1.1 Data Mining	1
1.2 R	2
1.3 Datasets	2
1.3.1 The Iris Dataset	2
1.3.2 The Bodyfat Dataset	3
2 Data Import and Export	5
2.1 Save and Load R Data	5
2.2 Import from and Export to .CSV Files	5
2.3 Import Data from SAS	6
2.4 Import/Export via ODBC	8
2.4.1 Read from Databases	8
2.4.2 Output to and Input from EXCEL Files	9
3 Data Exploration	11
3.1 Have a Look at Data	11
3.2 Explore Individual Variables	13
3.3 Explore Multiple Variables	16
3.4 More Explorations	20
3.5 Save Charts into Files	25
4 Decision Trees and Random Forest	27
4.1 Decision Trees with Package <i>party</i>	27
4.2 Decision Trees with Package <i>rpart</i>	31
4.3 Random Forest	36
5 Regression	41
5.1 Linear Regression	41
5.2 Logistic Regression	47
5.3 Generalized Linear Regression	48
5.4 Non-Linear Regression	50
6 Clustering	51
6.1 The k-Means Clustering	51
6.2 The k-Medoids Clustering	53

6.3	Hierarchical Clustering	56
6.4	Density-Based Clustering	57
7	Outlier Detection	63
7.1	Univariate Outlier Detection	63
7.2	Outlier Detection with LOF	66
7.3	Outlier Detection by Clustering	70
7.4	Outlier Detection from Time Series	72
7.5	Discussions	73
8	Time Series Analysis and Mining	75
8.1	Time Series Data in R	75
8.2	Time Series Decomposition	76
8.3	Time Series Forecasting	78
8.4	Time Series Clustering	78
8.4.1	Dynamic Time Warping	79
8.4.2	Synthetic Control Chart Time Series Data	79
8.4.3	Hierarchical Clustering with Euclidean Distance	80
8.4.4	Hierarchical Clustering with DTW Distance	82
8.5	Time Series Classification	83
8.5.1	Classification with Original Data	83
8.5.2	Classification with Extracted Features	84
8.5.3	k -NN Classification	86
8.6	Discussions	87
8.7	Further Readings	87
9	Association Rules	89
9.1	Basics of Association Rules	89
9.2	The Titanic Dataset	90
9.3	Association Rule Mining	92
9.4	Removing Redundancy	96
9.5	Interpreting Rules	98
9.6	Visualizing Association Rules	99
9.7	Discussions and Further Readings	103
10	Text Mining	105
10.1	Retrieving Text from Twitter	105
10.2	Transforming Text	106
10.3	Stemming Words	108
10.4	Building a Term-Document Matrix	110
10.5	Frequent Terms and Associations	111
10.6	Word Cloud	113
10.7	Clustering Words	114
10.8	Clustering Tweets	116

10.8.1	Clustering Tweets with the k -Means Algorithm	116
10.8.2	Clustering Tweets with the k -Medoids Algorithm	118
10.9	Packages, Further Readings, and Discussions	121
11	Social Network Analysis	123
11.1	Network of Terms	123
11.2	Network of Tweets	127
11.3	Two-Mode Network	132
11.4	Discussions and Further Readings	136
12	Case Study I: Analysis and Forecasting of House Price Indices	137
12.1	Importing HPI Data	137
12.2	Exploration of HPI Data	138
12.3	Trend and Seasonal Components of HPI	145
12.4	HPI Forecasting	147
12.5	The Estimated Price of a Property	149
12.6	Discussion	149
13	Case Study II: Customer Response Prediction and Profit Optimization	151
13.1	Introduction	151
13.2	The Data of KDD Cup 1998	151
13.3	Data Exploration	160
13.4	Training Decision Trees	166
13.5	Model Evaluation	170
13.6	Selecting the Best Tree	173
13.7	Scoring	176
13.8	Discussions and Conclusions	179
14	Case Study III: Predictive Modeling of Big Data with Limited Memory	181
14.1	Introduction	181
14.2	Methodology	182
14.3	Data and Variables	182
14.4	Random Forest	183
14.5	Memory Issue	185
14.6	Train Models on Sample Data	186
14.7	Build Models with Selected Variables	188
14.8	Scoring	194
14.9	Print Rules	201
14.9.1	Print Rules in Text	201
14.9.2	Print Rules for Scoring with SAS	205
14.10	Conclusions and Discussion	211

15 Online Resources	213
15.1 R Reference Cards	213
15.2 R	213
15.3 Data Mining	214
15.4 Data Mining with R	216
15.5 Classification/Prediction with R	216
15.6 Time Series Analysis with R	216
15.7 Association Rule Mining with R	216
15.8 Spatial Data Analysis with R	217
15.9 Text Mining with R	217
15.10 Social Network Analysis with R	217
15.11 Data Cleansing and Transformation with R	218
15.12 Big Data and Parallel Computing with R	218
 R Reference Card for Data Mining	 221
 Bibliography	 225
 General Index	 229
 Package Index	 231
 Function Index	 233

List of Figures

3.1	Histogram	15
3.2	Density	15
3.3	Pie Chart	16
3.4	Bar Chart	16
3.5	Boxplot	18
3.6	Scatter Plot	18
3.7	Scatter Plot with Jitter	19
3.8	A Matrix of Scatter Plots	19
3.9	3D Scatter Plot	20
3.10	Heat Map	21
3.11	Level Plot	22
3.12	Contour	22
3.13	3D Surface	23
3.14	Parallel Coordinates	23
3.15	Parallel Coordinates with Package <i>lattice</i>	24
3.16	Scatter Plot with Package <i>ggplot2</i>	24
4.1	Decision Tree	29
4.2	Decision Tree (Simple Style)	30
4.3	Decision Tree with Package <i>rpart</i>	34
4.4	Selected Decision Tree	35
4.5	Prediction Result	36
4.6	Error Rate of Random Forest	38
4.7	Variable Importance	39
4.8	Margin of Predictions	40
5.1	Australian CPIs in Year 2008 to 2010	42
5.2	Prediction with Linear Regression Model	45
5.3	A 3D Plot of the Fitted Model	46
5.4	Prediction of CPIs in 2011 with Linear Regression Model	47
5.5	Prediction with Generalized Linear Regression Model	50
6.1	Results of k -Means Clustering	53
6.2	Clustering with the k -medoids Algorithm—I	54
6.3	Clustering with the k -medoids Algorithm—II	55
6.4	Cluster Dendrogram	56
6.5	Density-Based Clustering—I	58
6.6	Density-Based Clustering—II	59
6.7	Density-Based Clustering—III	59
6.8	Prediction with Clustering Model	60

7.1	Univariate Outlier Detection with Boxplot	64
7.2	Outlier Detection—I	65
7.3	Outlier Detection—II	66
7.4	Density of Outlier Factors	67
7.5	Outliers in a Biplot of First Two Principal Components	68
7.6	Outliers in a Matrix of Scatter Plots	69
7.7	Outliers with k -Means Clustering	71
7.8	Outliers in Time Series Data	73
8.1	A Time Series of AirPassengers	76
8.2	Seasonal Component	77
8.3	Time Series Decomposition	77
8.4	Time Series Forecast	78
8.5	Alignment with Dynamic Time Warping	79
8.6	Six Classes in Synthetic Control Chart Time Series	80
8.7	Hierarchical Clustering with Euclidean Distance	81
8.8	Hierarchical Clustering with DTW Distance	82
8.9	Decision Tree	84
8.10	Decision Tree with DWT	86
9.1	A Scatter Plot of Association Rules	100
9.2	A Balloon Plot of Association Rules	100
9.3	A Graph of Association Rules	101
9.4	A Graph of Items	102
9.5	A Parallel Coordinates Plot of Association Rules	102
10.1	Frequent Terms	112
10.2	Word Cloud	114
10.3	Clustering of Words	115
10.4	Clusters of Tweets	120
11.1	A Network of Terms—I	125
11.2	A Network of Terms—II	126
11.3	Distribution of Degree	128
11.4	A Network of Tweets—I	129
11.5	A Network of Tweets—II	130
11.6	A Network of Tweets—III	131
11.7	A Two-Mode Network of Terms and Tweets—I	133
11.8	A Two-Mode Network of Terms and Tweets—II	135
12.1	HPIs in Canberra from Jan. 1990 to Jan. 2011	139
12.2	Monthly Increase of HPI	140
12.3	Monthly Increase Rate of HPI	141
12.4	A Bar Chart of Monthly HPI Increase Rate	142
12.5	Number of Months with Increased HPI	143
12.6	Yearly Average Increase Rates of HPI	143
12.7	Monthly Average Increase Rates of HPI	144
12.8	Distribution of HPI Increase Rate	144
12.9	Distribution of HPI Increase Rate per Year	145

12.10	Distribution of HPI Increase Rate per Month	145
12.11	Decomposition of HPI Data	146
12.12	Seasonal Components of HPI Data	146
12.13	HPI Forecasting—I	148
12.14	HPI Forecasting—II	149
13.1	A Data Mining Process	152
13.2	Distribution of Response	156
13.3	Box Plot of Donation Amount	156
13.4	Barplot of Donation Amount	157
13.5	Histograms of Numeric Variables	161
13.6	Boxplot of HIT	162
13.7	Distribution of Donation in Various Age Groups	163
13.8	Distribution of Donation in Various Age Groups	164
13.9	Scatter Plot	165
13.10	Mosaic Plots of Categorical Variables	166
13.11	A Decision Tree	169
13.12	Total Donation Collected (1000—400—4—10)	171
13.13	Total Donation Collected (9 runs)	172
13.14	Average Result of Nine Runs	173
13.15	Comparison of Different Parameter Settings—I	175
13.16	Comparison of Different Parameter Settings—II	175
13.17	Validation Result	178
14.1	Decision Tree	191
14.2	Test Result—I	192
14.3	Test Result—II	193
14.4	Test Result—III	194
14.5	Distribution of Scores	200

