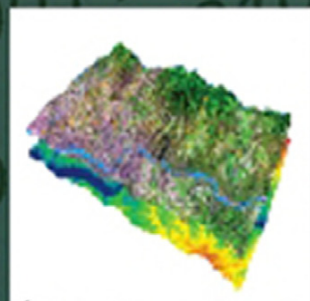




PETROLEUM INDUSTRY PRESS

# Data Mining and Knowledge Discovery for Geoscientists



Guangren Shi

# DATA MINING AND KNOWLEDGE DISCOVERY FOR GEOSCIENTISTS

---

This page intentionally left blank

# DATA MINING AND KNOWLEDGE DISCOVERY FOR GEOSCIENTISTS

---

GUANGREN SHI

*Professor of Mathematical Geology,  
Research Institute of Petroleum Exploration and Development,  
Beijing, China*



ELSEVIER

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

Elsevier

525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

225 Wyman Street, Waltham, MA 02451, USA

Copyright © 2014 Petroleum Industry Press. 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](mailto: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

Shi, Guangren.

Data mining and knowledge discovery for geoscientists/Guangren Shi, professor of mathematical geology, Research Institute of Petroleum Exploration and Development, Beijing, China. – First edition.  
pages cm

Includes bibliographical references.

ISBN 978-0-12-410437-2 (hardback)

1. Geology—Data processing. 2. Data mining. I. Title.

QE48.8.S54 2014

006.3'12—dc23

2013032294

#### British Library Cataloguing in Publication Data

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

For information on all Elsevier publications visit our web site at [store.elsevier.com](http://store.elsevier.com)

Printed and bound in USA

14 15 16 17 18 10 9 8 7 6 5 4 3 2 1

ISBN: 978-0-12-410437-2



Working together  
to grow libraries in  
developing countries

[www.elsevier.com](http://www.elsevier.com) • [www.bookaid.org](http://www.bookaid.org)

# Contents

---

## Preface vii

## 1 Introduction 1

- 1.1. Introduction to Data Mining 3
- 1.2. Data Systems Usable by Data Mining 7
- 1.3. Commonly Used Regression and Classification Algorithms 12
- 1.4. Data Mining System 16
- Exercises 20
- References 21

## 2 Probability and Statistics 23

- 2.1. Probability 25
- 2.2. Statistics 36
- Exercises 52
- References 52

## 3 Artificial Neural Networks 54

- 3.1. Methodology 56
- 3.2. Case Study 1: Integrated Evaluation of Oil and Gas-Trap Quality 74
- 3.3. Case Study 2: Fractures Prediction Using Conventional Well-Logging Data 80
- Exercises 86
- References 86

## 4 Support Vector Machines 87

- 4.1. Methodology 89
- 4.2. Case Study 1: Gas Layer Classification Based on Porosity, Permeability, and Gas Saturation 95
- 4.3. Case Study 2: Oil Layer Classification Based on Well-Logging Interpretation 101

- 4.4. Dimension-Reduction Procedure Using Machine Learning 105
- Exercises 109
- References 109

## 5 Decision Trees 111

- 5.1. Methodology 113
- 5.2. Case Study 1: Top Coal Caving Classification (Twenty-Nine Learning Samples) 125
- 5.3. Case Study 2: Top Coal Caving Classification (Twenty-Six Learning Samples and Three Prediction Samples) 131
- Exercises 138
- References 138

## 6 Bayesian Classification 139

- 6.1. Methodology 143
- 6.2. Case Study 1: Reservoir Classification in the Fuxin Uplift 163
- 6.3. Case Study 2: Reservoir Classification in the Baibao Oilfield 168
- 6.4. Case Study 3: Oil Layer Classification Based on Well-Logging Interpretation 173
- 6.5. Case Study 4: Integrated Evaluation of Oil and Gas Trap Quality 176
- 6.6. Case Study 5: Coal-Gas-Outburst Classification 180
- 6.7. Case Study 6: Top Coal Caving Classification (Twenty-Six Learning Samples and Three Prediction Samples) 186
- Exercises 190
- References 190

7 Cluster Analysis	191	9.3. Fractal Geometry	301
7.1. Methodology	194	9.4. Linear Programming	306
7.2. Case Study 1: Integrated Evaluation of Oil and Gas Trap Quality	207	Exercises	318
7.3. Case Study 2: Oil Layer Classification Based on Well-Logging Interpretation	215	References	319
7.4. Case Study 3: Coal-Gas-Outburst Classification	222	10 A Practical Software System of Data Mining and Knowledge Discovery for Geosciences	320
7.5. Case Study 4: Reservoir Classification in the Baibao Oilfield	229	10.1. Typical Case Study 1: Oil Layer Classification in the Keshang Formation	322
Exercises	237	10.2. Typical Case Study 2: Oil Layer Classification in the Lower H3 Formation	329
References	237	10.3. Typical Case Study 3: Oil Layer Classification in the Xiefengqiao Anticline	334
8 Kriging	238	10.4. A Practical System of Data Mining and Knowledge Discovery for Geosciences	334
8.1. Preprocessing	241	Exercises	340
8.2. Experimental Variogram	245	References	340
8.3. Optimal Fitting of Experimental Variogram	250	Appendix 1: Table of Unit Conversion	341
8.4. Cross-Validation of Kriging	259	Appendix 2: Answers to Exercises	345
8.5. Applications of Kriging	267	Index	361
8.6. Summary and Conclusions	274		
Exercises	274		
References	274		
9 Other Soft Computing Algorithms for Geosciences	275		
9.1. Fuzzy Mathematics	279		
9.2. Gray Systems	282		

# Preface

---

This book is an aggregation of principles, methods, codes, and applications for the data mining and knowledge discovery in geosciences based on the author's studies over the past 17 years.

In the past 20 years, the field of data mining has seen an enormous success in terms of both wide-ranging applications and scientific methodologies. *Data mining* is the computerized process of extracting previously unknown and important actionable information and knowledge from large databases. Such knowledge can then be used to make crucial decisions by incorporating individuals' intuition and experience so as to objectively generate for decision makers informed options that might otherwise go undiscovered. So, data mining is also called *knowledge discovery in database*, and it has been widely applied in many fields of economics, science, and technology. However, data mining applications to geosciences are still at an initial stage, partly due to the multidisciplinary nature and complexity of geosciences and partly due to the fact that many new methods in data mining require time and well-tested case studies in geosciences.

Facing the challenges of large amounts of geosciences databases, geoscientists can use database management systems to conduct conventional applications (such as queries, searches, and simple statistical analysis), but they cannot obtain the available knowledge inherent in data by such methods, leading to a paradoxical scenario of "rich data but poor knowledge." The true solution is to apply data mining techniques in

geosciences databases and modify such techniques to suit practical applications in geosciences. This book, *Data Mining and Knowledge Discovery for Geoscientists*, is a timely attempt to summarize the latest developments in data mining for geosciences.

This book introduces some successful applications of data mining in geosciences in recent years for knowledge discovery in geosciences. It systematically introduces to geoscientists the widely used algorithms and discusses their basic principles, conditions of applications, and diversity of case studies as well as describing what algorithm may be suitable for a specific application.

This book focuses on eight categories of algorithm: (1) probability and statistics, (2) artificial neural networks, (3) support vector machines, (4) decision trees, (5) Bayesian classification, (6) cluster analysis, (7) Kriging method, and (8) other soft computing algorithms, including fuzzy mathematics, gray systems, fractal geometry, and linear programming.

This consists of 22 algorithms: probability density function, Monte Carlo method, least-squares method constructing linear function, least-squares constructing exponent function, least-squares constructing polynomial, multiple regression analysis, back-propagation neural network, the classification of support vector machine, the regression of support vector machine, ID3 decision trees, C4.5 decision trees, naïve Bayesian, Bayesian discrimination, Bayesian successive discrimination, Q-mode cluster analysis, R-mode cluster analysis, Kriging,



fuzzy integrated decision, gray prediction, gray integrated decision, fractal geometry, and linear programming. For each algorithm, its applying ranges and conditions, basic principles, calculating method, calculation flowchart, and one or more detailed case studies are discussed. The book contains 41 case studies, 38 of which are in the area of geosciences. In each case study, for classification and regression algorithms, the solution accuracy comparison and algorithm selection have been made. Finally, a practical system of data mining and knowledge discovery for

geosciences is presented. Moreover, this book also provides some exercises in each chapter; answers to all exercises are provided in a special appendix. Therefore, this book is dedicated to two kinds of people: (1) researchers and programmers in data mining, scientists and engineers in geosciences, and university students and lecturers in geosciences; and (2) scientists and engineers in computer science and information technology and university students and lecturers in information-related subjects such as database management.

# Introduction

## OUTLINE

<b>1.1. Introduction to Data Mining</b>	<b>3</b>	<b>1.2. Data Systems Usable by Data Mining</b>	<b>7</b>
1.1.1. <i>Motivity of Data Mining</i>	3	1.2.1. <i>Databases</i>	7
1.1.2. <i>Objectives and Scope of Data Mining</i>	3	1.2.1.1. Database types	7
1.1.2.1. Generalization	4	1.2.1.2. Data Properties	8
1.1.2.2. Association	4	1.2.1.3. Development Phases	8
1.1.2.3. Classification and Clustering	4	1.2.1.4. Commonly Used Databases	10
1.1.2.4. Prediction	5	1.2.2. <i>Data Warehousing</i>	10
1.1.2.5. Deviation	5	1.2.2.1. Data Storage	10
1.1.3. <i>Classification of Data Mining Systems</i>	5	1.2.2.2. Construction Step	11
1.1.3.1. To Classify According to the Mined DB Type	5	1.2.3. <i>Data Banks</i>	11
1.1.3.2. To Classify According to the Mined Knowledge Type	6	<b>1.3. Commonly Used Regression and Classification Algorithms</b>	<b>12</b>
1.1.3.3. To Classify According to the Available Techniques Type	6	1.3.1. <i>Linear and Nonlinear Algorithms</i>	13
1.1.3.4. To Classify According to the Application	6	1.3.2. <i>Error Analysis of Calculation Results</i>	14
1.1.4. <i>Major Issues in Data Mining for Geosciences</i>	6	1.3.3. <i>Differences between Regression and Classification Algorithms</i>	15
		1.3.4. <i>Nonlinearity of a Studied Problem</i>	15
		1.3.5. <i>Solution Accuracy of Studied Problem</i>	15
		<b>1.4. Data Mining System</b>	<b>16</b>
		1.4.1. <i>System Functions</i>	16
		1.4.2. <i>System Flowcharts</i>	18