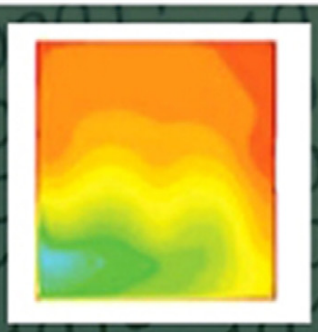
ELSEVIER

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



**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



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

# Contents

# 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.

CHAPTER

# 1

# Introduction

OUTLINE




*Data Mining and Knowledge Discovery for Geoscientists*
http://dx.doi.org/10.1016/B978-0-12-410437-2.00001-1