

DATA MINING ALGORITHMS



Alpha
Science

Rajan Chattamvelli

DATA MINING ALGORITHMS

Rajan Chattamvelli



Alpha Science International Ltd.
Oxford, U.K.

DATA MINING ALGORITHMS

424 pgs. | 40 figs. | 87 tbls.

Rajan Chattamvelli

Associate Professor

Department of Information Technology

Periyar Maniammai University

Thanjavur, Tamil Nadu

India 613 403

Copyright © 2011

ALPHA SCIENCE INTERNATIONAL LTD.

7200 The Quorum, Oxford Business Park North

Garsington Road, Oxford OX4 2JZ, U.K.

www.alphasci.com

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 prior written permission of the publisher.

Printed from the camera-ready copy provided by the Author.

ISBN 978-1-84265-684-6

Printed in India

DATA MINING ALGORITHMS

Preface

This is an intermediate level textbook on data mining with emphasis on algorithms and applications. It brings under one roof the multitude of algorithms available for common data mining tasks. Numerical examples illustrate most of the algorithms. Pseudo code for some of the algorithms is given to benefit the students and researchers. Links to software programs are given at the end of each chapter for the benefit of the readers.

This book can be used for advanced undergraduate or graduate level courses in data mining, machine learning, and soft-computing. Some of the chapters can also be used for courses on statistics, econometrics and management; as well as for competitive examinations. This is also an ideal book for researchers in various fields. A brief overview of the chapters is given below.

Chapter 1 introduces data types and scales of measurement. An understanding of this concept is essential for data miners and researchers. This occupies more than 50% of the material in this chapter. It also discusses the ‘Date’ data type in depth (§1.4.7, pp.1-12), and gives a table of date formats used in various countries (pp.1-13). This is useful when the data for loading into datawarehouses are collected from various countries, or when data are captured using web forms that are filled online by people in different countries. The chapter then goes on to discuss data warehouses and data marts. This is followed by a thorough discussion on supervised and unsupervised learning. An up-to-date discussion on data discretisation algorithms appears in §1.8 in page 1-19. Various steps in data mining, and popular data mining approaches are then discussed. This chapter ends with a few practical applications from various fields.

Chapter 2 introduces basic concepts in probability in an intuitive way. Important results are briefly summarised, followed by a thorough discussion of basic concepts in Statistics. Measures of location, dispersion and skewness are discussed. Data outliers and their detecting techniques are briefly explained. Data transformation techniques are thoroughly explained with lots of examples (pp.2-27). This is not only useful to data miners, but also to engineers, statisticians, and scientific programmers. Linear regression and correlation are introduced (pp.2-32), and some simple expressions for covariance are derived. New recursive algorithms for sample variance and covariance are also derived, and exemplified (pp.2-40→2-42). A brief discussion of Monte Carlo methods (pp.2-48) and contingency tables (pp.2-49) can be found at the end of the chapter. This chapter will be useful to researchers from various fields, and even to undergraduate and graduate students in statistics, econometrics, engineering, medical sciences and management.

Chapter 3 introduces decision trees (DT). The concept of classification is introduced in section 3.2 (pp.3-8), followed by the most popular measures for node splitting (§3.3, pp.3-13). Popular tree induction algorithms are discussed in detail, and their features are compared. DT induction algorithms are discussed in §3.4, and a comparison table is given in pp.3-25. The chapter ends with a list of software for decision tree modeling.

The chapter 4 introduces association rules, which is not as popular as other data mining models due to its applicability in focused fields. Topics covered include association rule measures, cross-purchase and sequence purchase analysis, activity indicators, etc. Special association rules like negative associations, sparse associations, rare associations, temporal associations are discussed. Pareto analysis and paired comparison analysis are discussed. This is followed by a thorough discussion of ARM algorithms and their extensions. The FP-tree algorithm, which is the most popular, is extensively discussed with numerical examples. Dynamic FP-growth, and modified FP-growth algorithms are discussed. Various association rule mining algorithms are then compared. A few practical applications of association rules are then given. This is followed by a list of software for association rules.

Chapter 5 is on web mining. The first few sections introduce the basics needed to understand the rest of the chapter. Web content mining and web structure mining are introduced. This is followed by an extensive discussion of web structure mining. The Original Page Rank Algorithm (OPRA) of Brin & Page is introduced §5.5, and its statistical distribution is derived. The HITS and OPRA algorithms are numerically illustrated. The OPRA is generalised in the next section to obtain a variety of useful generalised page rank algorithms. These include Noise Removed Page Rank Algorithm (NoRPRA), Alpha Page Rank Algorithm (APRA), Filtered Page Rank Algorithm (FiPRA), Weighted Page Rank Algorithm (WePRA), and Hybrid Page Rank Algorithm (HyPRA). A discussion of TrustRank algorithm follows next. The rest of the chapter discusses semantic web mining, text mining, image mining and table mining.

Chapter 6 on support vector machines (SVM) is more mathematical than the other chapters. Some knowledge in geometry (vectors), matrices and linear algebra, differential calculus and quadratic programming is needed to understand the entire chapter. Those who are not familiar with these topics can still benefit from the first few sub-sections and the application section. It starts with structural risk minimisation principle and various solution techniques. Linear separability and hyperplane classifiers are discussed. Binary SVM (with 2 classes) is discussed next, followed by confidence in classification. A new “canonical hyperplane theorem” can be found in page 6-10. Lagrangian formulation of the classical SVM is then described and the dual SVM is obtained. Soft-margin and weighted SVM are described next, followed by multi-class SVM, ν -SVM and LP-SVM. The Sequential Minimal Optimisation (SMO) is discussed at length next. This is followed by a thorough discussion of LS-SVM, which is numerically illustrated. Other topics include the I-SVM, support vector regression (SVR), and non-linear SVM. An SVM summary table appears in page 6-42, followed by a table of SVM classifiers. A thorough discussion on SVM vs Statistical Classifiers appears next. This is followed by variable selection methods using SVM. Popular SVM software appears at the end.

Chapter 7 introduces vector space models (VSM). An extension of VSM called latent semantic analysis (LSA) maps the input data to a reduced rank feature space using truncated singular value decomposition (SVD) principle. Latent semantic indexing (LSI) is an adaptation of LSA to information retrieval (IR). It originated with text retrieval, but has been extended to other types of data. Topics covered include the SVD algorithm, forming the LSI query, and query execution details. Applications of LSI to information retrieval and clustering are given at the end. An automatic labeling technique to uniquely identify clusters found by an LSA algorithm is also discussed. Some sections of this chapter are also mathematical, requiring basic knowledge in geometry, matrices and linear algebra. Software for LSI appears at the end.

The last chapter introduces spatial data warehousing and mining. Common problems in spatial data mining are described. Characteristics of spatial and geospatial data are discussed next. Spatial windowing techniques, spatial map overlay etc are discussed in §8.2. A discussion on spatial data transformations appears in §8.3. Geographical Information Systems (GIS) and geo-spatial operators are introduced in §8.4. Spatial and classical data mining are then compared in §8.6. Classical and spatial autocorrelations are thoroughly introduced in §8.7. Spatial indexing techniques are described next. Several spatial algorithms like spatial association rules, spatial clustering and classification, spatial trend detection and interpolation are discussed. The chapter ends with a list of software for SDM.

Chapter 2 can be used for undergraduate courses in statistics, econometrics, and management. Other chapters may be used as ‘supplementary reading’ for various courses – chapters 2,3 for decision support courses; 1,5,7 for Information Retrieval courses; 2, 3, 8 for Econometrics courses; 1, 2, 3, 5 and 8 for management. Professionals and practitioners in various fields can also use the book for self-study. The pre-requisites include one course in Statistics (chapters 2,3,8), some knowledge in matrices and linear algebra (chapter 4–7), calculus (chapter 6) and data structures (chapter 3, 4, 8). Some basic concepts from geometry are needed to understand chapters 6–8. All other data mining topics like clustering, genetic algorithms, neural networks, text mining, data visualisation and OLAP, data warehouses, etc are discussed in a companion volume “Data Mining Methods (2nd ed.)”.

A carefully selected set of exercises has been provided to benefit the students and self-study professionals. Answers and hints are provided for selected exercises. Any suggestions and comments for improvement are welcome. All suggestions should be sent to dmmbook@gmail.com. Up-to-date errata will be made available upon request.

Rajan Chattamvelli
Tanjore, Tamil Nadu.

Table of Contents

1. INTRODUCTION TO DATA MINING	1-1
1.1 Introduction	1-1
1.2 Data and Measurement	1-1
1.2.1 Meaning of Information	1-2
1.2.2 Computer Data Types	1-3
1.3 Data Categories	1-3
1.4 Categorical and Quantitative Data	1-3
1.4.1 The NOIR Scale of Measurement	1-4
1.4.2 Nominal Scale	1-4
Coding of Nominal Variables	1-6
Binary Data	1-7
Coding of Binary Variables	1-7
Symmetric vs Asymmetric Binary Variables	1-7
Nominal Dichotomisation	1-8
Ternary Data	1-9
1.4.3 Ordinal Scale	1-10
1.4.4 Types of Ordinal Data	1-10
Ordinal data transformation	1-11
1.4.5 Allowed Operations on Ordinal Data	1-11
1.4.6 Interval Scale	1-12
1.4.7 The Date Data Type	1-12
Allowed Operations on Interval Data	1-14
Interval Data Transformations	1-14
1.4.8 Ratio Scale	1-15
Operations on Ratio Data	1-15
Choosing the Correct Scale	1-15
1.4.9 Conversion Between Scales	1-15
1.5 The Extended Scales of Measurement	1-16
1.5.1 Nonstandard Data	1-17
1.6 The STO Scale	1-17
1.7 Supervised and Unsupervised Learning	1-17
1.7.1 Supervised Models	1-18
1.7.2 Unsupervised Models	1-19
1.8 Numeric Data Discretisation	1-19
1.8.1 A Taxonomy of Data Discretisation	1-20
1.8.2 Equal Interval Binning (EIB)	1-21

1.8.3	Equal Frequency Binning (EFB)	1-22
	Error in Discretisation	1-23
1.9	Databases and Data Warehouses	1-24
1.9.1	Data Warehouses	1-25
1.10	Data Mining	1-25
1.10.1	Exploratory Data Analysis vs Data Mining	1-25
1.11	Steps in Data Mining	1-26
1.11.1	Data Mining Approaches	1-27
1.11.2	Applications	1-28
1.11.3	Summary	1-28
1.12	Exercises	1-28
	References	1-33
2.	PROBABILITY AND STATISTICS	2-1
2.1	Introduction	2-1
2.2	Probability	2-3
2.2.1	Different Ways to Express Probability	2-3
2.2.2	A Notation for Probability	2-4
2.2.3	Methods of Counting	2-6
	Independence of Events	2-7
2.2.4	Rules of Probability	2-7
	Probability Model	2-9
	Entropy vs Probability	2-9
2.3	Venn Diagrams	2-10
2.3.1	De'Morgan's Laws	2-10
2.4	Bayes Theorem	2-11
2.4.1	Bayes Theorem for Conditional Probability	2-12
	Odds-Likelihood Ratio Form of Bayes Theorem	2-13
	Product Rule for Conditional Probability	2-13
2.4.2	Bayes Classification Rule	2-13
	Rule of Expected Utility	2-13
2.5	Mathematical Expectation	2-14
2.6	Statistics	2-14
2.6.1	Population vs Sample	2-14
2.6.2	Parameter vs Statistic	2-15
2.7	Measures of Location	2-16
2.7.1	Mean, Median and Mode	2-16
	Weighted Mean	2-18
	Advantages of Mean	2-19
2.7.2	Median	2-19
	Advantages of Median	2-20
2.7.3	Mode	2-20
	Advantages of Mode	2-21
2.7.4	Geometric Mean	2-22
2.7.5	Harmonic Mean	2-22
2.7.6	Which Mean is Most Appropriate?	2-23
2.8	Measures of Dispersion	2-23
2.8.1	Range	2-24

2.8.2	Inter-Quartile Range	2-24
2.8.3	Mean Absolute Deviation	2-24
2.8.4	Variance	2-25
2.9	Outliers in Data	2-25
2.9.1	Spatial vs Temporal Outliers	2-26
2.9.2	Graphical Detection of Outliers	2-26
2.10	Data Transformations	2-27
2.10.1	Change of Origin	2-27
2.10.2	Change of Scale	2-27
2.10.3	Change of Origin and Scale	2-28
2.10.4	Min-max Transformation	2-29
2.10.5	Transformation to Symmetric Range	2-30
2.10.6	Standard Normalisation	2-32
2.10.7	Nonlinear Transformations	2-32
2.11	Regression Basics	2-32
2.11.1	Scatterplots and Regression	2-33
Advantages of Scatter Plots	2-33	
2.11.2	Simple Linear Regression	2-35
2.11.3	Ordinary Least Squares (OLS)	2-36
2.12	Sample Covariance	2-37
2.12.1	Recursive Algorithm for Sample Covariance	2-40
2.12.2	Weighted Least Squares (WLS)	2-45
2.12.3	Correlation Coefficient	2-46
2.12.4	From Scatterplot to Correlation	2-46
Interpretation of Correlation Coefficient	2-46	
2.12.5	Multivariate Data	2-47
2.13	Multiple Linear Regression (MLR)	2-47
2.14	Monte Carlo Methods	2-48
Components of Monte Carlo Simulation	2-49	
2.15	Contingency Tables	2-49
2.16	Exercises	2-50
References	2-54	
3.	DECISION TREES	3-1
3.1	Trees	3-1
3.1.1	Binary Trees	3-2
Drawing Trees	3-2	
3.2	Decision Trees	3-3
3.2.1	Chance and Terminal Nodes	3-3
3.2.2	Advantages of Decision Trees	3-4
3.2.3	Disadvantages of Decision Trees	3-7
3.2.4	Classification	3-8
3.2.5	Production Rules	3-9
3.2.6	Building a DT	3-12
3.3	Measures for Node Splitting	3-13
3.3.1	Gini's Index Measure	3-13
3.3.2	Minimum Classification Error Measure	3-14
3.3.3	Shannon's Entropy Measure	3-14

3.3.4	Gain and Impurity	3-16
3.4	Induction Algorithms	3-17
3.4.1	ID3 Algorithm	3-17
3.4.2	C4.5 Algorithm	3-18
3.4.3	Extended Tabular Method to Build a Decision Tree	3-19
3.4.4	CHI-squared Automatic Interaction Detector (CHAID)	3-23
3.4.5	Classification and Regression Tree (CART)	3-24
3.4.6	Misclassification Errors	3-24
3.5	Pruning Decision Trees	3-26
3.5.1	Pre and Post Pruning	3-26
3.6	Fuzzy Decision Trees	3-27
Decision Tables		3-27
3.7	Applications	3-28
Fraud Detection		3-28
Pruning ATM Decision Tree		3-31
3.8	Software for Decision Trees	3-32
3.9	Exercises	3-33
References		3-36
4.	ASSOCIATION RULES	4-1
4.1	Meaning of Association Rules	4-1
4.1.1	Motivation for Association Rules	4-2
4.1.2	Uni-Directional and Bi-Directional Associations	4-3
4.1.3	Antecedent and Consequent	4-4
4.1.4	Categorical Variables	4-6
4.2	Association Rule Measures	4-6
4.2.1	Support	4-6
4.2.2	Confidence	4-7
4.2.3	Lift	4-9
4.2.4	Cover	4-9
4.3	Association Rule Mining	4-10
4.3.1	Cross-purchase Analysis	4-11
4.3.2	Sequence-purchase Analysis	4-11
4.3.3	Activity Indicators	4-12
4.4	Special Association Rules	4-13
4.4.1	Negative Associations	4-13
4.4.2	Negative Association Rules	4-15
4.4.3	Sparse Association Rules	4-16
4.4.4	Rare Associations	4-17
4.4.5	Temporal Association Rules	4-18
4.4.6	Pareto Analysis	4-19
4.4.7	The Inverse Pareto Principle	4-19
4.4.8	Paired Comparisons Analysis	4-20
4.4.9	Fuzzy Association Rules	4-20
4.4.10	Plan Mining	4-21
4.5	Generalisations of Association Rules (GAR)	4-21
4.6	Extended Association Rules	4-22
4.6.1	Multi-Level Association Rules (MLAR)	4-22

4.6.2	Multi-Dimensional Association Rules (MDAR)	4-23
4.6.3	Constrained Association Rules	4-23
4.6.4	Rule Constraints in Association Rule Mining	4-23
4.6.5	Weighted Association Rule Mining (WARM)	4-24
4.7	Algorithms for Association Rules	4-24
4.7.1	The Apriori Principle	4-24
4.7.2	The AIS Algorithm	4-25
4.7.3	Apriori Algorithm	4-25
4.7.4	Variants of Apriori Algorithm	4-29
4.8	The FP-Growth Algorithm (FPGA)	4-31
4.8.1	FP-Tree	4-32
4.8.2	Advantages of FP-Growth Algorithm	4-34
4.8.3	The FP-Tree Construction	4-34
4.8.4	Conditional Pattern Bases (CPB)	4-36
4.8.5	Forming the CPT	4-38
4.9	Dynamic FP-Growth Algorithm	4-41
4.10	Modified FP-Growth Algorithm	4-42
4.10.1	Other Algorithms	4-45
4.11	Applications	4-45
4.11.1	Purchase Domain Application	4-45
4.11.2	Diagnosis	4-46
4.11.3	Inventory Arrangement	4-48
4.11.4	Fraud Detection	4-48
4.12	Software for Association Rules	4-49
4.13	Exercises	4-50
	References	4-54
5.	WEB MINING	5-1
5.1	Web Pages	5-1
5.2	Web Sites	5-2
5.2.1	The PAO Graph of Web Sites	5-4
5.3	Search Engines	5-5
5.3.1	Indexers	5-6
5.3.2	Information Extraction	5-6
5.3.3	Linguistic Search Engines	5-6
5.4	Web Mining	5-7
5.4.1	Advantages of Web Mining	5-7
5.5	Implementing Web Mining	5-8
5.5.1	Web Content Mining (WCM)	5-9
5.5.2	Web Usage Mining (WUM)	5-10
5.5.3	Web User-Quality Mining (WUQM)	5-10
5.6	Web Structure Mining (WSM)	5-11
5.6.1	Link Mining	5-11
5.6.2	The Page Views	5-12
5.7	Measures for Web Structure Mining	5-13
5.7.1	Heuristic Ranking Algorithms	5-14
5.7.2	Hubs and Authorities (H & A) Algorithm	5-15
5.7.3	PageRank Algorithm (PRA)	5-17

5.7.4	The Original Page Rank Algorithm (OPRA)	5-18
5.7.5	Quality of Inlinks	5-18
5.7.6	Computing the Page-Rank	5-19
5.7.7	Drawback of PageRank Algorithm	5-19
5.8	Generalised PageRank Algorithms	5-22
5.8.1	PageRank Equations	5-22
5.8.2	Noise-Removed Page Rank Algorithm (NoRPRA)	5-23
5.8.3	Alpha Page Rank Algorithm (APRA)	5-23
5.8.4	Weighted Page Rank Algorithm (WePRA)	5-24
5.8.5	Filtered Page Rank Algorithm (FiPRA)	5-24
5.8.6	Hybrid Page Rank Algorithm (HyPRA)	5-25
5.8.7	TrustRank Algorithm	5-25
5.8.8	Link Categorisation	5-25
5.8.9	Link Stepping	5-26
5.8.10	Links Analysis	5-27
5.8.11	Web Query Mining (WQM)	5-28
5.8.12	Query Performance Measures	5-28
	The F-score	5-29
5.9	Semantic Web Mining	5-29
5.9.1	Metadata Mining	5-30
5.9.2	Multilingual Web Mining	5-31
5.9.3	Web Personalisers	5-31
5.10	Applications	5-31
	Spam-Mail Classification	5-32
	Web-page Clustering	5-32
	Web Marketing	5-33
	Miscellaneous Applications	5-33
5.11	Software for Web and Text Mining	5-34
5.12	Exercises	5-34
	References	5-36
6.	SUPPORT VECTOR MACHINES	6-1
6.1	Introduction	6-1
6.1.1	Structural Risk Minimisation Principle	6-3
6.1.2	Solution Techniques	6-3
6.1.3	Linear Separability	6-3
6.1.4	Hyperplane Classifiers	6-4
6.2	Binary SVM	6-5
6.2.1	Binary SVM Classifier	6-6
6.2.2	Binary SVM Margin	6-6
	Confidence in Classification	6-7
6.2.3	Simple SVM (SSVM)	6-8
6.2.4	ρ -SVM	6-11
6.3	Lagrangian Formulation	6-12
6.3.1	Dual SVM Formulation	6-13
	Estimating the Intercept Term	6-13
6.3.2	Properties of Dual SVM	6-14
	Dual of ρ -SVM	6-15

6.4	Weighted SVM (W-SVM)	6-17
	Classifying the Unclassifiable	6-18
6.4.1	Overlapping Classes	6-18
6.5	Soft-Margin SVM (SM-SVM)	6-19
6.5.1	Weighted Soft Margin SVM (WSM-SVM)	6-21
6.6	Multi-class SVM (MC-SVM)	6-21
6.6.1	Pair-wise SSVM (One-versus-One [OVO])	6-21
6.6.2	One-versus-All (OVA) SVM	6-22
6.7	ν -SVM	6-23
6.8	LP-SVM	6-24
6.8.1	L_1L_2 SVM	6-24
6.9	Sequential Minimal Optimisation (SMO)	6-25
	Computing Intercept Term b	6-29
6.9.1	Advantages of SMO Algorithm	6-31
6.9.2	Pruning	6-31
6.10	Kernels	6-33
6.10.1	Properties of Kernels	6-34
6.10.2	Mercer's Theorem	6-35
6.11	Least Squares SVM (LS-SVM)	6-35
6.11.1	LS-SVM Formulation	6-35
6.12	Incremental SVM (I-SVM)	6-38
6.12.1	I-SVM Formulation	6-39
6.12.2	I-SVM Training	6-40
6.13	Nonlinear SVM (NL-SVM)	6-41
6.13.1	Other Kernel Algorithms	6-43
6.14	Support Vector Regression (SVR)	6-43
6.15	SVM vs Statistical Classifiers	6-45
6.15.1	Boundary Dense and Boundary Sparse Problems	6-45
6.15.2	Duality Theorem and SVM	6-46
6.16	Variable Selection Using SVM	6-47
6.17	Applications of SVM	6-49
6.17.1	Medical Application	6-49
6.17.2	Text Categorisation	6-51
6.18	SVM Software	6-53
6.19	Exercises	6-54
	References	6-58
7.	LATENT SEMANTIC INDEXING	7-1
7.1	Vector Space Models	7-1
7.1.1	Term-by-Document Matrix	7-2
7.1.2	Textual IR	7-3
7.1.3	Geometric Interpretation	7-4
7.2	Latent Semantic Analysis	7-5
7.2.1	Steps in LSA	7-5
7.2.2	Characteristics of LSA	7-6
7.2.3	Advantages of LSA	7-6
7.2.4	Disadvantages of LSA	7-9
7.3	Singular Value Decomposition	7-10

7.4	LSI Query	7-12
7.4.1	Query Processing	7-12
7.5	Applications of LSI	7-15
7.5.1	LSI in Information Retrieval	7-15
7.5.2	Latent Semantic Clustering (LSC)	7-19
	Labeling Semantic Clusters Found	7-21
7.6	Software for LSI	7-22
7.7	Exercises	7-22
	References	7-24
8.	SPATIAL DATA MINING	8-1
8.1	Introduction	8-1
8.1.1	Spatial Data Mining	8-2
8.1.2	Problems in Spatial Data Mining	8-3
8.1.3	Aims of Spatial Data Mining	8-4
8.2	Spatial Data	8-4
	Geocoding	8-4
8.2.1	Geospatial Data and Maps	8-9
8.2.2	Spatial Windowing	8-11
8.2.3	Spatial Map Overlay (SMO)	8-12
8.2.4	Spatial Data Storage	8-17
8.2.5	Spatial And Non-spatial Database (SAND)	8-19
8.2.6	SOLAP	8-20
8.2.7	Spatio-Temporal Databases (STD)	8-21
8.3	Spatial Data Transformation	8-22
8.4	Geographic Information Systems	8-23
	Mercator projection	8-23
8.4.1	Geospatial Operators	8-24
8.5	Spatial Statistics vs Spatial Data Mining	8-25
8.6	Spatial vs Classical Data Mining	8-25
8.6.1	Time Series vs Temporal Data Mining	8-26
8.6.2	Taxonomy of Spatial Operations	8-27
8.6.3	Spatial Joins	8-30
8.6.4	Spatial Convolutions	8-30
8.7	Autocorrelations	8-31
8.7.1	Classical vs Spatial Autocorrelations	8-31
8.7.2	Spatial Autocorrelations	8-32
	Moran Coefficient (MC)	8-33
	Geary Coefficient (GC)	8-37
8.7.3	Spatial Access Methods (SAM)	8-37
8.8	Spatial Indexing	8-38
8.8.1	Spatial B-Trees	8-39
8.8.2	Spatial Structured Query Languages (SSQL)	8-39
8.8.3	Spatio-Temporal Query Languages (ST-SQL)	8-40
8.9	Categorisation of SDM Tasks	8-41
8.9.1	Spatial Search and Optimisation	8-41
8.9.2	Distributed Spatial Data Mining (DSDM)	8-42
8.10	Knowledge Discovery in Spatial Databases	8-42

8.10.1	Spatial Outlier Detection Algorithms (SODA)	8-42
8.10.2	Spatial Association Rule Mining (SARM)	8-44
8.10.3	Spatial Classification	8-45
8.10.4	Spatial Clustering	8-46
8.10.5	Spatial Trend Detection	8-46
8.10.6	Spatial Interpolation	8-47
	Kriging	8-48
8.10.7	Miscellaneous Topics	8-48
	Spatial Data Repositories	8-49
8.10.8	Software for Spatial and Geo-Spatial Data Mining	8-49
8.11	Exercises	8-49
	References	8-52
Appendix I – Solutions to Selected Exercises		A-1
Appendix II – List of Acronyms		A-16
Author Index		AI-1
Subject Index		SI-1

List of Algorithms

1.1	Equal Interval Binning Algorithm	1-21
1.2	Equal Interval Binning Algorithm – Nested For Loop	1-21
1.3	Equal Frequency Binning Algorithm	1-23
2.1	Algorithm for Contrast Stretching	2-29
2.2	Algorithm for Asymmetric to Symmetric Range Transformation	2-31
2.3	Recursive Algorithm for Sample Covariance	2-44
3.1	ID3 Algorithm	3-18
4.1	Apriori Algorithm for Positive Association Rules	4-27
4.2	Algorithm RuleGen For Rule Generation	4-28
4.3	Streamlined Apriori Algorithm for Positive Association Rules	4-30
4.4	High-level Pseudocode of the FP-Growth Algorithm	4-42
4.5	FP-Growth Algorithm	4-43
4.6	The FP-Tree Mining Algorithm	4-44
5.1	Algorithm for Hubs and Authorities	5-15
5.2	Algorithm for Computing the PageRank of Interlinked Pages	5-20
6.1	Canonical Hyperplane Algorithm	6-11
6.2	Sequential Minimal Optimisation Algorithm – Main program	6-29
6.3	Sequential Minimal Optimisation Algorithm – Induct values	6-30
6.4	SMO Algorithm continued.....Jointly Optimise Lagrange Multipliers	6-32
6.5	High-level Pseudocode for Variable Selection in Support Vector Machines	6-47
6.6	Pseudocode of Support Vector Variable Selection	6-48
6.7	Algorithm for Primal Variable Selection in Support Vector Machines	6-50
7.8	Pseudocode of the SVD Algorithm	7-12
7.9	Algorithm for Term (resp. Document) Clustering Using LSC	7-21
8.1	Algorithm to Convert Fractional Longitude/Latitude to Minutes and Seconds	8-8
8.2	Algorithm to Convert Milli-Arc-Second to Degree, and Fractional Minute	8-9
8.3	Algorithm for Manual Map Overlay	8-12
8.4	Algorithm to Find Surface Distance Using Longitude and Latitudes	8-29
8.5	Algorithm to Fit Autoregressive Models	8-31