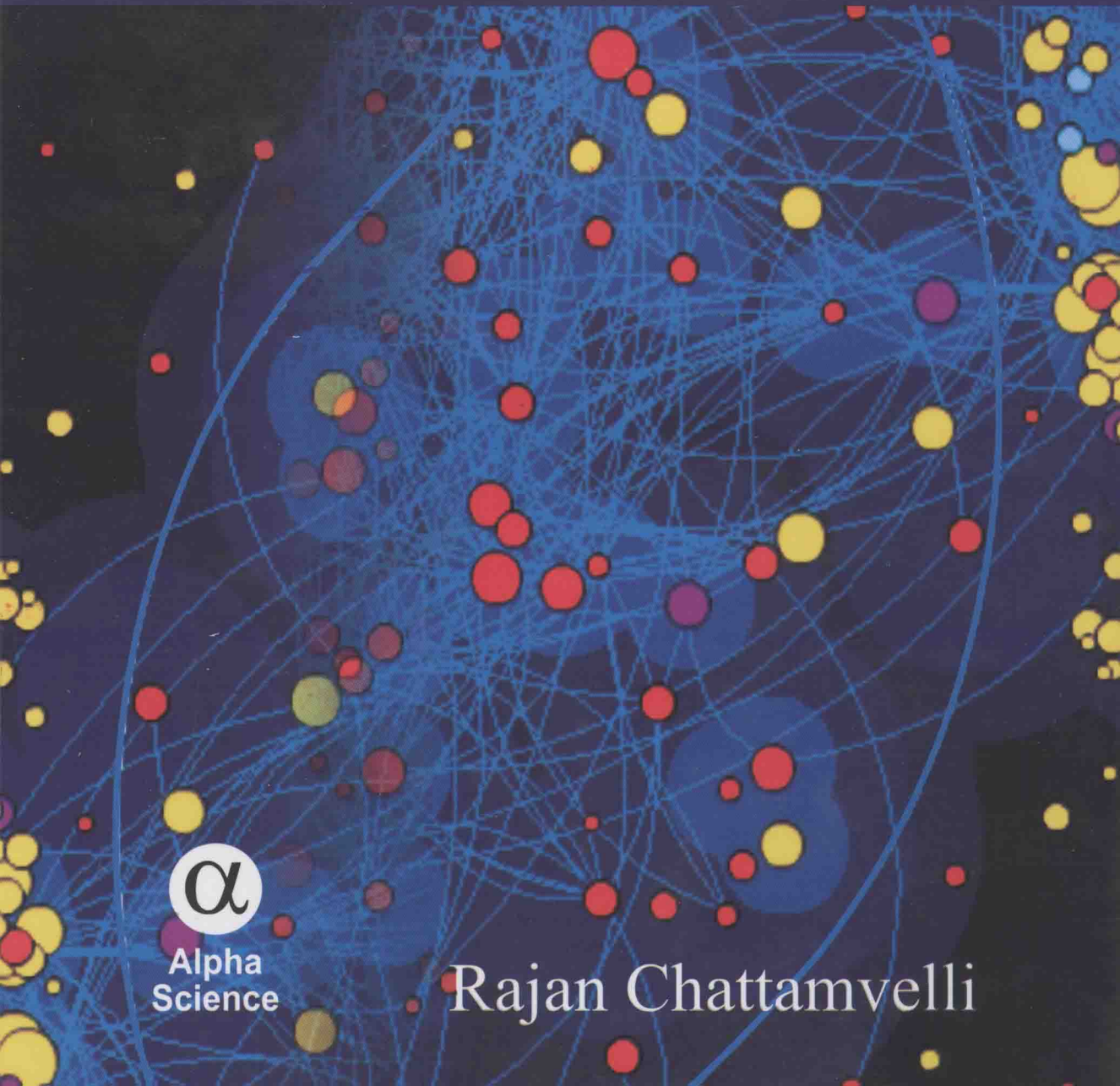
# DATA MINING ALGORITHMS



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



ALPHA SCIENCE INTERNATIONAL LTD.
7200 The Quorum, Oxford Business Park North
Garsington Road, Oxford OX4 2JZ, U.K.

**www.alphasci.com**



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, $\nu-$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 ($2^{\text{nd}}$ 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

# List of Algorithms