Studies in Computational Intelligence 190

K.R. Venugopal
K.G. Srinivasa
L.M. Patnaik

Soft Computing for Data Mining Applications

Springer

K.R. Venugopal
K.G. Srinivasa
L.M. Patnaik

Soft Computing for Data Mining Applications

Springer

Dr. K.R. Venugopal
Dean, Faculty of Engineering
University Visvesvaraya College of
Engineering
Bangalore University
Bangalore 560001
Karnataka
India

Prof. L.M. Patnaik
Professor, Vice Chancellor
Defence Institute of
Advanced Technology
Deemed University
Girinagar, Pune 411025
India

Dr. K.G. Srinivasa
Assistant Professor,
Department of Computer Science and
Engineering
M.S. Ramaiah Institute of Technology
MSRIT Post,
Bangalore 560054
Karnataka
India

ISBN 978-3-642-00192-5 e-ISBN 978-3-642-00193-2

DOI 10.1007/978-3-642-00193-2

Studies in Computational Intelligence ISSN 1860949X

Library of Congress Control Number: 2008944107

Typeset & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India.

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Foreword

The authors have consolidated their research work in this volume titled Soft Computing for Data Mining Applications. The monograph gives an insight into the research in the fields of Data Mining in combination with Soft Computing methodologies. In these days, the data continues to grow exponentially. Much of the data is implicitly or explicitly imprecise. Database discovery seeks to discover noteworthy, unrecognized associations between the data items in the existing database. The potential of discovery comes from the realization that alternate contexts may reveal additional valuable information. The rate at which the data is stored is growing at a phenomenal rate. As a result, traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analyzing this vast collection of data. Several domains where large volumes of data are stored in centralized or distributed databases includes applications like in electronic commerce, bioinformatics, computer security, Web intelligence, intelligent learning database systems, finance, marketing, healthcare, telecommunications, and other fields.

Efficient tools and algorithms for knowledge discovery in large data sets have been devised during the recent years. These methods exploit the capability of computers to search huge amounts of data in a fast and effective manner. However, the data to be analyzed is imprecise and afflicted with uncertainty. In the case of heterogeneous data sources such as text and video, the data might moreover be ambiguous and partly conflicting. Besides, patterns and relationships of interest are usually approximate. Thus, in order to make the information mining process more robust it requires tolerance toward imprecision, uncertainty and exceptions.

With the importance of soft computing applied in data mining applications in recent years, this monograph gives a valuable research directions in the field of specialization. As the authors are well known writers in the field of Computer Science and Engineering, the book presents state of the art technology in data mining. The book is very useful to researchers in the field of data mining.

Bangalore,
November 2008

N.R. Shetty
President, ISTE, India

Preface

In today's digital age, there is huge amount of data generated everyday. Deriving meaningful information from this data is a huge problem for humans. Therefore, techniques such as data mining whose primary objective is to unearth hithero unknown relationship from data becomes important. The application of such techniques varies from business areas (Stock Market Prediction, Content Based Image Retrieval), Proteomics (Motif Discovery) to Internet (XML Data Mining, Web Personalization). The traditional computational techniques find it difficult to accomplish this task of Knowledge Discovery in Databases (KDD). Soft computing techniques like Genetic Algorithms, Artificial Neural Networks, Fuzzy Logic, Rough Sets and Support Vector Machines when used in combination is found to be more effective. Therefore, soft computing algorithms are used to accomplish data mining across different applications.

Chapter one presents introduction to the book. Chapter two gives details of self adaptive genetic algorithms. An iterative merge based genetic algorithms for data mining applications is given in chapter three. Dynamic association rule mining using genetic algorithms is described in chapter four. An evolutionary approach for XML data mining is presented in chapter five. Chapter six, gives a neural network based relevance feedback algorithm for content based image retrieval. An hybrid algorithm for predicting share values is addressed in chapter seven. The usage of rough sets and genetic algorithms for data mining based query processing is discussed in chapter eight. An effective web access sequencing algorithm using hashing techniques for better web reorganization is presented in chapter nine. An efficient data structure for personalizing the Google search results is mentioned in chapter ten. Classification based clustering algorithms using naive Bayesian probabilistic models are discussed in chapter eleven. The effective usage of simulated annealing and genetic algorithms for mining top-k ranked webpages from Google is presented in chapter twelve. The concept of mining bioXML databases is introduced in chapter thirteen. Chapter fourteen and fifteen discusses algorithms for DNA compression. An efficient algorithm for motif discovery in protein sequences is presented in chapter sixteen. Finally, matching techniques for genome sequences and genetic algorithms for motif discovery are given in chapter seventeen and eighteen respectively.

The authors appreciate the suggestions from the readers and users of this book. Kindly communicate the errors, if any, to the following email address: venugopalkr@gmail.com.

Bangalore,
November 2008

K.R. Venugopal
K.G. Srinivasa
L.M. Patnaik

Acknowledgements

We wish to place on record our deep debt of gratitude to Shri M C Jayadeva, who has been a constant source of inspiration. His gentle encouragement have been the key for the growth and success in our career. We are indebted to Prof. K Venkatagiri Gowda for his inspiration, encouragement and guidance throughout our lives. We thank Prof. N R Shetty, President, ISTE and Former Vice Chancellor, Bangalore University, Bangalore for his foreword to this book. We owe debt of gratitude to Sri K Narahari, Sri V Nagaraj, Prof. S Lakshmana Reddy, Prof. K Mallikarjuna Chetty, Prof. H N Shivashankar, Prof. P Sreenivas Kumar, Prof. Kamala Krithivasan, Prof. C Sivarama Murthy, Prof. T Basavaraju, Prof. M Channa Reddy, Prof. N Srinivasan, Prof. M Venkatachalappa for encouraging us to bring out this book in the present form. We sincerely thank Sri K P Jayarama Reddy, T G Girikumar, P Palani, M G Muniyappa for their support in the preparation of this book.

We are grateful to Justice M Rama Jois, Sri N Krishnappa for their encouragement. We express our gratitude to Sri Y K Raghavendra Rao, Sri P R Ananda Rao, Justice T Venkataswamy, Prof. V Y Somayajulu, Sri Sreedhar Sagar, Sri N Nagabhusan, Sri Prabhakar Bhat, Prof. K V Acharya, Prof. Khajampadi Subramanya Bhat, Sri Dinesh Kamath, Sri D M Ravindra, Sri Jagadeesh Karanath, Sri N Thippeswamy, Sri Sudhir, Sri V Manjunath, Sri N Dinesh Hegde, Sri Nagendra Prasad, Sri Sripad, Sri K Thyagaraj, Smt. Savithri Venkatagiri Gowda, Smt. Karthyayini V and Smt. Rukmini T, our well wishers for inspiring us to write this book.

We thank Prof. K S Ramanatha, Prof. K Rajanikanth, V K Ananthashayana and T V Suresh Kumar for their support. We thank Smt. P Deepa Shenoy, Sri K B Raja, Sri K Suresh Babu, Smt. J Triveni, Smt. S H Manjula, Smt. D N Sujatha, Sri Prakash G L, Smt. Vibha Lakshmikantha, Sri K Girish, Smt. Anita Kanavalli, Smt. Alice Abraham, Smt. Shaila K, for their suggestions and support in bringing out this book.

We are indebted to Tejaswi Venugopal, T Shivaprakash, T Krishnaprasad and Lakshmi Priya K for their help. Special thanks to Nalini L and Hemalatha for their invaluable time and neat desktop composition of the book.

About the Authors

K.R. Venugopal is Principal and Dean, Faculty of Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore. He obtained his Bachelor of Technology from University Visvesvaraya College of Engineering in 1979. He received his Masters degree in Computer Science and Automation from Indian Institute of Science Bangalore He was awarded Ph.D. in Economics from Bangalore University and Ph.D. in Computer Science from Indian Institute of Technology, Madras. He has a distinguished academic career and has degrees in Electronics, Economics, Law, Business Finance, Public Relations, Communications, Industrial Relations, Computer Science and Journalism. He has authored and edited twenty seven books on Computer Science and Economics, which include Petrodollar and the World Economy, Programming with Pascal, Programming with FORTRAN, Programming with C, Microprocessor Programming, Mastering C++ etc. He has been serving as the Professor and Chairman, Department of Computer Science and Engineering, UVCE. He has over two hundred research papers in refereed International Journals and Conferences to his credit. His research interests include computer networks, parallel and distributed systems and database systems.

K.G. Srinivasa obtained his a Ph.D. in Computer Science and Engineering from Bangalore University. Currently he is working as an Assistant Professor in the Department of Computer Science and Engineering, M S Ramaiah Institute of Technology, Bangalore. He received Bachelors and Masters degree in Computer Science and Engineering from the Bangalore University in the year 2000 and 2002 respectively. He is a member of IEEE, IETE, and ISTE. He has authored more than fifty research papers in refereed International Journals and Conferences. His research interests are Soft Computing, Data Mining and Bioinformatics.

L.M. Patnaik is Vice Chancellor of Defence Institute of Advanced Studies, Pune, India. He was the Professor since 1986 with the Department of

Computer Science and Automation, Indian Institute of Science, Bangalore. During the past 35 years of his service at the Institute. He has over 400 research publications in in refereed International Journals and Conference Proceedings. He is a Fellow of all the four leading Science and Engineering Academies in India; Fellow of the IEEE and the Academy of Science for the Developing World. He has received twenty national and international awards; notable among them is the IEEE Technical Achievement Award for his significant contributions to high performance computing and soft computing. His areas of research interest have been parallel and distributed computing, mobile computing, CAD for VLSI circuits, soft computing, and computational neuroscience.

Contents

Acronyms

GA	Genetic Algorithms
ANN	Artificial Neural Networks
AI	Artificial Intelligence
SVM	Support Vector Machines
KDD	Knowledge Discovery in Databases
OLAP	On-Line Analytical Processing
MIQ	Machine Intelligence Quotient
FL	Fuzzy Logic
RS	Rough Sets
XML	eXtended Markup Language
HTML	Hyper Text Markup Language
SQL	Structured Query Language
PCA	Principal Component Analysis
SDI	Selective Dissemination of Information
SOM	Self Organizing Map
CBIR	Content Based Image Retrieval
WWW	World Wide Web
DNA	Deoxyribo Nucleic Acid
IGA	Island model Genetic Algorithms
SGA	Simple Genetic Algorithms
PID	Pima Indian Diabetes
Wisc	Wisconsin Breast Cancer Database
Hep	Hepatitis Database
Ion	Ionosphere Database
LVQ	Learning Vector Quantization
BPNN	Backpropagation Neural Network
RBF	Radial Basis Function
ITI	Incremental Decision Tree Induction
LMDT	Linear Machine Decision Tree
DTD	Document Type Definition
MFI	Most Frequently used Index

LFI	Less Frequently used Index
hvi	Hierarchical Vector Identification
UIC	User Interest Categories
KNN	k Nearest Neighborhood
DMQL	Data Mining Query Languages
TSP	Travelling Salesman Problem
MAD	Mean Absolute Deviation
SSE	Sum of Squared Error
MSE	Mean Squared Error
RMSE	Root Mean Squared Error
MAPE	Mean Absolute Percentage Error
STI	Shape Texture Intensity
HIS	Hue, Intensity and Saturation
DCT	Discrete Cosine Transform
PWM	Position Weight Matrix
PSSM	Position Specific Scoring Matrix
PRDM	Pairwise Relative Distance Matrix
DSSP	Secondary Structure of Proteins
LSI	Latent Semantic Indexing
GIS	Geographical Information Systems
CAD	Computer Aided Design
FS	Free Search
BGA	Breeder Genetic Algorithm
STIRF	Shape, Texture, Intensity-distribution features with Relevance Feedback