The Top Ten Algorithms in Data Mining

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

SERIES EDITOR

Vipin Kumar
University of Minnesota
Department of Computer Science and Engineering
Minneapolis, Minnesota, U.S.A

AIMS AND SCOPE

This series aims to capture new developments and applications in data mining and knowledge discovery, while summarizing the computational tools and techniques useful in data analysis. This series encourages the integration of mathematical, statistical, and computational methods and techniques through the publication of a broad range of textbooks, reference works, and handbooks. The inclusion of concrete examples and applications is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of data mining and knowledge discovery methods and applications, modeling, algorithms, theory and foundations, data and knowledge visualization, data mining systems and tools, and privacy and security issues.

PUBLISHED TITLES

UNDERSTANDING COMPLEX DATASETS: Data Mining with Matrix Decompositions
David Skillicorn

COMPUTATIONAL METHODS OF FEATURE SELECTION
Huan Liu and Hiroshi Motoda

CONSTRAINED CLUSTERING: Advances in Algorithms, Theory, and Applications
Sugato Basu, Ian Davidson, and Kiri L. Wagstaff

KNOWLEDGE DISCOVERY FOR COUNTERTERRORISM AND LAW ENFORCEMENT
David Skillicorn

MULTIMEDIA DATA MINING: A Systematic Introduction to Concepts and Theory
Zhongfei Zhang and Ruofei Zhang

NEXT GENERATION OF DATA MINING
Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar

DATA MINING FOR DESIGN AND MARKETING
Yukio Ohsawa and Katsutoshi Yada

GEOGRAPHIC DATA MINING AND KNOWLEDGE DISCOVERY, Second Edition
Harvey J. Miller and Jiawei Han

THE TOP TEN ALGORITHMS IN DATA MINING
Xindong Wu and Vipin Kumar

Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

The Top Ten Algorithms in Data Mining

Edited by
Xindong Wu
Vipin Kumar

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-1-4200-8964-6 (Hardcover)

© 2009 by Taylor & Francis Group, LLC

Contents

Preface

In an effort to identify some of the most influential algorithms that have been widely used in the data mining community, the IEEE International Conference on Data Mining (ICDM, http://www.cs.uvm.edu/∼icdm/) identified the top 10 algorithms in data mining for presentation at ICDM '06 in Hong Kong. This book presents these top 10 data mining algorithms: C4.5, *k*-Means, SVM, Apriori, EM, PageRank, AdaBoost, *k*NN, Naïve Bayes, and CART.

As the first step in the identification process, in September 2006 we invited the ACM KDD Innovation Award and IEEE ICDM Research Contributions Award winners to each nominate up to 10 best-known algorithms in data mining. All except one in this distinguished set of award winners responded to our invitation. We asked each nomination to provide the following information: (a) the algorithm name, (b) a brief justification, and (c) a representative publication reference. We also advised that each nominated algorithm should have been widely cited and used by other researchers in the field, and the nominations from each nominator as a group should have a reasonable representation of the different areas in data mining.

After the nominations in step 1, we verified each nomination for its citations on Google Scholar in late October 2006, and removed those nominations that did not have at least 50 citations. All remaining (18) nominations were then organized in 10 topics: association analysis, classification, clustering, statistical learning, bagging and boosting, sequential patterns, integrated mining, rough sets, link mining, and graph mining. For some of these 18 algorithms, such as *k*-means, the representative publication was not necessarily the original paper that introduced the algorithm, but a recent paper that highlights the importance of the technique. These representative publications are available at the ICDM Web site (http://www.cs.uvm.edu/∼icdm/algorithms/CandidateList.shtml).

In the third step of the identification process, we had a wider involvement of the research community. We invited the Program Committee members of KDD-06 (the 2006 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining), ICDM '06 (the 2006 IEEE International Conference on Data Mining), and SDM '06 (the 2006 SIAM International Conference on Data Mining), as well as the ACM KDD Innovation Award and IEEE ICDM Research Contributions Award winners to each vote for up to 10 well-known algorithms from the 18-algorithm candidate list. The voting results of this step were presented at the ICDM '06 panel on Top 10 Algorithms in Data Mining.

At the ICDM '06 panel of December 21, 2006, we also took an open vote with all 145 attendees on the top 10 algorithms from the above 18-algorithm candidate list,

and the top 10 algorithms from this open vote were the same as the voting results from the above third step. The three-hour panel was organized as the last session of the ICDM '06 conference, in parallel with seven paper presentation sessions of the Web Intelligence (WI '06) and Intelligent Agent Technology (IAT '06) conferences at the same location, and attracted 145 participants.

After ICDM '06, we invited the original authors and some of the panel presenters of these 10 algorithms to write a journal article to provide a description of each algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm. The journal article was published in January 2008 in *Knowledge and Information Systems* [1]. This book expands upon this journal article, with a common structure for each chapter on each algorithm, in terms of algorithm description, available software, illustrative examples and applications, advanced topics, and exercises.

Each book chapter was reviewed by two independent reviewers and one of the two book editors. Some chapters went through a major revision based on this review before their final acceptance.

We hope the identification of the top 10 algorithms can promote data mining to wider real-world applications, and inspire more researchers in data mining to further explore these 10 algorithms, including their impact and new research issues. These 10 algorithms cover classification, clustering, statistical learning, association analysis, and link mining, which are all among the most important topics in data mining research and development, as well as for curriculum design for related data mining, machine learning, and artificial intelligence courses.

Acknowledgments

The initiative of identifying the top 10 data mining algorithms started in May 2006 out of a discussion between Dr. Jiannong Cao in the Department of Computing at the Hong Kong Polytechnic University (PolyU) and Dr. Xindong Wu, when Dr. Wu was giving a seminar on 10 Challenging Problems in Data Mining Research [2] at PolyU. Dr. Wu and Dr. Vipin Kumar continued this discussion at KDD-06 in August 2006 with various people, and received very enthusiastic support.

Naila Elliott in the Department of Computer Science and Engineering at the University of Minnesota collected and compiled the algorithm nominations and voting results in the three-step identification process. Yan Zhang in the Department of Computer Science at the University of Vermont converted the 10 section submissions in different formats into the same LaTeX format, which was a time-consuming process.

Xindong Wu and Vipin Kumar
September 15, 2008

References

[1] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand, and Dan Steinberg, Top 10 algorithms in data mining, *Knowledge and Information Systems*, 14(2008), 1: 1–37.

[2] Qiang Yang and Xindong Wu (Contributors: Pedro Domingos, Charles Elkan, Johannes Gehrke, Jiawei Han, David Heckerman, Daniel Keim, Jiming Liu, David Madigan, Gregory Piatetsky-Shapiro, Vijay V. Raghavan, Rajeev Rastogi, Salvatore J. Stolfo, Alexander Tuzhilin, and Benjamin W. Wah), 10 challenging problems in data mining research, *International Journal of Information Technology & Decision Making*, 5, 4(2006), 597–604.

About the Authors

Xindong Wu is a professor and the chair of the Computer Science Department at the University of Vermont, United States. He holds a PhD in Artificial Intelligence from the University of Edinburgh, Britain. His research interests include data mining, knowledge-based systems, and Web information exploration. He has published over 170 referred papers in these areas in various journals and conferences, including IEEE TKDE, TPAMI, ACM TOIS, DMKD, KAIS, IJCAI, AAAI, ICML, KDD, ICDM, and WWW, as well as 18 books and conference proceedings. He won the IEEE ICTAI-2005 Best Paper Award and the IEEE ICDM-2007 Best Theory/Algorithms Paper Runner Up Award.

Dr. Wu is the editor-in-chief of the *IEEE Transactions on Knowledge and Data Engineering (TKDE*, by the IEEE Computer Society), the founder and current Steering Committee Chair of the *IEEE International Conference on Data Mining (ICDM)*, the founder and current honorary editor-in-chief of Knowledge and Information Systems (KAIS, by Springer), the founding chair (2002–2006) of the IEEE Computer Society Technical Committee on Intelligent Informatics (TCII), and a series editor of the Springer Book Series on Advanced Information and Knowledge Processing (AI&KP). He served as program committee chair for ICDM '03 (the 2003 IEEE International Conference on Data Mining) and program committee cochair for KDD-07 (the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining). He is the 2004 ACM SIGKDD Service Award winner, the 2006 IEEE ICDM Outstanding Service Award winner, and a 2005 chair professor in the Changjiang (or Yangtze River) Scholars Programme at the Hefei University of Technology sponsored by the Ministry of Education of China and the Li Ka Shing Foundation. He has been an invited/keynote speaker at numerous international conferences including NSF-NGDM'07, PAKDD-07, IEEE EDOC'06, IEEE ICTAI'04, IEEE/WIC/ACM WI'04/IAT'04, SEKE 2002, and PADD-97.

Vipin Kumar is currently William Norris professor and head of the Computer Science and Engineering Department at the University of Minnesota. He received BE degrees in electronics and communication engineering from Indian Institute of Technology, Roorkee (formerly, University of Roorkee), India, in 1977, ME degree in electronics engineering from Philips International Institute, Eindhoven, Netherlands, in 1979, and PhD in computer science from University of Maryland, College Park, in 1982. Kumar's current research interests include data mining, bioinformatics, and high-performance computing. His research has resulted in the development of the concept of isoefficiency metric for evaluating the scalability of parallel algorithms, as well as highly efficient parallel algorithms and software for sparse matrix factorization

(PSPASES) and graph partitioning (METIS, ParMetis, hMetis). He has authored over 200 research articles, and has coedited or coauthored 9 books, including widely used textbooks *Introduction to Parallel Computing* and *Introduction to Data Mining*, both published by Addison-Wesley. Kumar has served as chair/cochair for many conferences/workshops in the area of data mining and parallel computing, including *IEEE International Conference on Data Mining* (2002), *International Parallel and Distributed Processing Symposium* (2001), and *SIAM International Conference on Data Mining* (2001). Kumar serves as cochair of the steering committee of the *SIAM International Conference on Data Mining*, and is a member of the steering committee of the *IEEE International Conference on Data Mining* and the *IEEE International Conference on Bioinformatics and Biomedicine*. Kumar is a founding coeditor-in-chief of *Journal of Statistical Analysis and Data Mining*, editor-in-chief of *IEEE Intelligent Informatics Bulletin*, and editor of *Data Mining and Knowledge Discovery Book Series*, published by CRC Press/Chapman Hall. Kumar also serves or has served on the editorial boards of *Data Mining and Knowledge Discovery*, *Knowledge and Information Systems*, *IEEE Computational Intelligence Bulletin*, *Annual Review of Intelligent Informatics*, *Parallel Computing*, the *Journal of Parallel and Distributed Computing*, *IEEE Transactions of Data and Knowledge Engineering* (1993–1997), *IEEE Concurrency* (1997–2000), and *IEEE Parallel and Distributed Technology* (1995–1997). He is a fellow of the ACM, IEEE, and AAAS, and a member of SIAM. Kumar received the 2005 IEEE Computer Society's Technical Achievement award for contributions to the design and analysis of parallel algorithms, graph-partitioning, and data mining.

Contributors

Songcan Chen, *Nanjing University of Aeronautics and Astronautics, Nanjing, China*

Joydeep Ghosh, *University of Texas at Austin, Austin, TX*

David J. Hand, *Imperial College, London, UK*

Alexander Liu, *University of Texas at Austin, Austin, TX*

Bing Liu, *University of Illinois at Chicago, Chicago, IL*

Geoffrey J. McLachlan, *University of Queensland, Brisbane, Australia*

Hiroshi Motoda, *ISIR, Osaka University and AFOSR/AOARD, Air Force Research Laboratory, Japan*

Shu-Kay Ng, *Griffith University, Meadowbrook, Australia*

Kouzou Ohara, *ISIR, Osaka University, Japan*

Naren Ramakrishnan, *Virginia Tech, Blacksburg, VA*

Michael Steinbach, *University of Minnesota, Minneapolis, MN*

Dan Steinberg, *Salford Systems, San Diego, CA*

Pang-Ning Tan, *Michigan State University, East Lansing, MI*

Hui Xue, *Nanjing University of Aeronautics and Astronautics, Nanjing, China*

Qiang Yang, *Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong*

Philip S. Yu, *University of Illinois at Chicago, Chicago, IL*

Yang Yu, *Nanjing University, Nanjing, China*

Zhi-Hua Zhou, *Nanjing University, Nanjing, China*