

Outlier Ensembles

Charu C. Aggarwal · Saket Sathe

Outlier Ensembles

An Introduction

 Springer

Charu C. Aggarwal
IBM T. J. Watson Research Center
Yorktown Heights, NY
USA

Saket Sathe
IBM T. J. Watson Research Center
Yorktown Heights, NY
USA

ISBN 978-3-319-54764-0 ISBN 978-3-319-54765-7 (eBook)
DOI 10.1007/978-3-319-54765-7

Library of Congress Control Number: 2017934615

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Charu C. Aggarwal dedicates this book to his wife Lata and his daughter Sayani.

Saket Sathe dedicates this book to his wife Aishwarya and his son Devansh.

Preface

Talent wins games, but teamwork and intelligence wins championships.

Michael Jordan

Ensemble analysis is a widely used class of meta-algorithms for many data mining problems such as classification and clustering. Numerous ensemble-based algorithms have been proposed in the literature for these problems. Compared to the clustering and classification problems, ensemble analysis has been studied in a limited way in the context of outlier detection. It is only in recent years that the problem of outlier ensembles has been recognized formally, and some techniques for these problems have been proposed. However, the material in the field has not been formally recognized, and it is expected that this book will play a significant role in creating a structured exposition of the topic. This book discusses a variety of methods for outlier ensembles and organizes them by the specific principles with which accuracy improvements are achieved. In addition, a discussion is provided on the techniques with which such methods can be made more effective. A formal classification of these methods has been provided, and the circumstances in which they work well are discussed. A discussion is also provided on how outlier ensembles relate (both theoretically and practically) to the ensemble techniques used commonly for other data mining problems such as classification. The similarities and (subtle) differences in the ensemble techniques for the classification and outlier detection problems are discussed. These subtle differences do impact the design of ensemble algorithms for the latter problem.

Ensemble techniques are increasingly finding their way into the curricula of many data mining courses. This book can be used for such courses. Many illustrative examples and exercises are provided in order to facilitate classroom teaching. A familiarity is assumed to the outlier detection problem and also to the generic problem of ensemble analysis in classification. This is because many of the ensemble methods discussed in this book are adaptations from their counterparts in the classification domain. Some techniques discussed in this book, such as

wagging, randomized feature weighting, and geometric subsampling, provide new insights that are not available elsewhere. We have also provided an analysis of the performance of various types of base detectors and their relative effectiveness. This book combines a textbook-style discussion of older topics with new insights. Therefore, we believe that the book will also be of interest to researchers and practitioners for leveraging ensemble methods into optimal algorithmic design.

Yorktown Heights, NY, USA
January 2017

Charu C. Aggarwal
Saket Sathe

Acknowledgements

We would like to thank our families for their support during the writing of this book. Charu C. Aggarwal would like to thank his wife Lata and his daughter Sayani. Saket Sathé would like to thank his wife Aishwarya for her love and support. He would like to especially thank his four-year-old son Devansh for his affection and playfulness.

We are grateful to the support of our management at IBM. In particular, we would like to thank Nagui Halim and Deepak Turaga for their support during the writing of this book.

We would like to thank our numerous collaborators whose insights have helped make this book a success. Charu Aggarwal would like to thank Tarek F. Abdelzaher, Jing Gao, Quanquan Gu, Manish Gupta, Jiawei Han, Alexander Hinneburg, Thomas Huang, Nan Li, Huan Liu, Ruoming Jin, Daniel Keim, Arijit Khan, Latifur Khan, Mohammad M. Masud, Jian Pei, Magda Procopiuc, Guojun Qi, Chandan Reddy, Jaideep Srivastava, Karthik Subbian, Yizhou Sun, Jiliang Tang, Min-Hsuan Tsai, Haixun Wang, Jianyong Wang, Min Wang, Suhang Wang, Joel Wolf, Xifeng Yan, Philip Yu, Mohammed Zaki, ChengXiang Zhai, and Peixiang Zhao. Charu Aggarwal would also like to thank Lata Aggarwal for her help in some of the diagrams drawn using Powerpoint in the book. Saket Sathé would like to thank Karl Aberer, Timos Sellis, Dipanjan Chakraborty, Arun Vishwanath, Hoyoung Jeung, Sue Ann Chen, and Tian Guo for their collaborations over the years. We also thank Leman Akoglu for several general discussions on ensemble methods.

Contents

1	An Introduction to Outlier Ensembles	1
1.1	Introduction	1
1.1.1	Motivations for Ensemble Methods in Outlier Analysis	3
1.1.2	Common Settings for Existing Ensemble Methods	4
1.1.3	Types of Ensemble Methods	6
1.1.4	Overview of Outlier Ensemble Design	7
1.2	Categorization by Component Independence	8
1.2.1	Sequential Ensembles	9
1.2.2	Independent Ensembles	11
1.3	Categorization by Constituent Components	12
1.3.1	Model-Centered Ensembles	12
1.3.2	Data-Centered Ensembles	14
1.3.3	Discussion of Categorization Schemes	16
1.4	Categorization by Theoretical Approach	17
1.4.1	Variance Reduction in Outlier Ensembles	18
1.4.2	Bias Reduction in Outlier Ensembles	18
1.5	Defining Combination Functions	19
1.5.1	Normalization Issues	19
1.5.2	Combining Scores from Different Models	20
1.6	Research Overview and Book Organization	23
1.6.1	Overview of Book	27
1.7	Conclusions and Discussion	30
	References	31
2	Theory of Outlier Ensembles	35
2.1	Introduction	35
2.2	The Bias-Variance Trade-Off for Outlier Detection	37
2.2.1	Relationship of Ensemble Analysis to Bias-Variance Trade-Off	42
2.2.2	Out-of-Sample Issues	43

- 2.2.3 Understanding How Ensemble Analysis Works 44
- 2.2.4 Data-Centric View Versus Model-Centric View 50
- 2.3 Examples and Applications of the Bias-Variance Tradeoff. 58
 - 2.3.1 Bagging and Subsampling. 59
 - 2.3.2 Feature Bagging 60
 - 2.3.3 Boosting 61
- 2.4 Experimental Illustration of Bias-Variance Theory 61
 - 2.4.1 Understanding the Effects of Ensembles
on Data-Centric Bias and Variance 62
 - 2.4.2 Experimental Examples of Bias-Variance
Decomposition 68
- 2.5 Conclusions 72
- References. 73
- 3 Variance Reduction in Outlier Ensembles. 75**
 - 3.1 Introduction 75
 - 3.2 Motivations for Basic Variance Reduction Framework. 78
 - 3.3 Variance Reduction Is Not a Panacea. 83
 - 3.3.1 When Does Data-Centric Variance Reduction Help?. 84
 - 3.3.2 When Does Model-Centric Variance Reduction Help? 91
 - 3.3.3 The Subtle Differences Between AUCs and MSEs 93
 - 3.4 Variance Reduction Methods 93
 - 3.4.1 Feature Bagging (FB) for High-Dimensional Outlier
Detection. 94
 - 3.4.2 Rotated Bagging (RB). 99
 - 3.4.3 Projected Clustering and Subspace Histograms 100
 - 3.4.4 The Point-Wise Bagging and Subsampling Class
of Methods 107
 - 3.4.5 Wagging (WAG). 130
 - 3.4.6 Data-Centric and Model-Centric Perturbation 131
 - 3.4.7 Parameter-Centric Ensembles 131
 - 3.4.8 Explicit Randomization of Base Models 132
 - 3.5 Some New Techniques for Variance Reduction 134
 - 3.5.1 Geometric Subsampling (GS) 134
 - 3.5.2 Randomized Feature Weighting (RFW) 136
 - 3.6 Forcing Stability by Reducing Impact of Abnormal Detector
Executions 137
 - 3.6.1 Performance Analysis of Trimmed Combination
Methods 140
 - 3.6.2 Discussion of Commonly Used Combination Methods 143
 - 3.7 Performance Analysis of Methods 145
 - 3.7.1 Data Set Descriptions 145
 - 3.7.2 Comparison of Variance Reduction Methods 147

- 3.8 Conclusions 157
- References. 158
- 4 Bias Reduction in Outlier Ensembles: The Guessing Game. 163**
 - 4.1 Introduction 163
 - 4.2 Bias Reduction in Classification and Outlier Detection. 165
 - 4.2.1 Boosting 166
 - 4.2.2 Training Data Pruning 167
 - 4.2.3 Model Pruning 168
 - 4.2.4 Model Weighting 169
 - 4.2.5 Differences Between Classification and Outlier
Detection. 170
 - 4.3 Training Data Pruning 171
 - 4.3.1 Deterministic Pruning 171
 - 4.3.2 Fixed Bias Sampling 172
 - 4.3.3 Variable Bias Sampling. 174
 - 4.4 Model Pruning 175
 - 4.4.1 Implicit Model Pruning in Subspace Outlier Detection 178
 - 4.4.2 Revisiting Pruning by Trimming 178
 - 4.4.3 Model Weighting 180
 - 4.5 Supervised Bias Reduction with Unsupervised Feature
Engineering 181
 - 4.6 Bias Reduction by Human Intervention 182
 - 4.7 Conclusions 184
 - References. 184
- 5 Model Combination Methods for Outlier Ensembles 187**
 - 5.1 Introduction 187
 - 5.2 Impact of Outlier Evaluation Measures. 190
 - 5.3 Score Normalization Issues. 193
 - 5.4 Model Combination for Variance Reduction. 195
 - 5.5 Model Combination for Bias Reduction 196
 - 5.5.1 A Simple Example 198
 - 5.5.2 Sequential Combination Methods 199
 - 5.6 Combining Bias and Variance Reduction 200
 - 5.6.1 Factorized Consensus 201
 - 5.7 Using Mild Supervision in Model Combination 203
 - 5.8 Conclusions and Summary 204
 - References. 204
- 6 Which Outlier Detection Algorithm Should I Use?. 207**
 - 6.1 Introduction 207
 - 6.2 A Review of Classical Distance-Based Detectors 212
 - 6.2.1 Exact k -Nearest Neighbor Detector. 213
 - 6.2.2 Average k -Nearest Neighbor Detector. 214

6.2.3	An Analysis of Bagged and Subsampled 1-Nearest Neighbor Detectors	214
6.2.4	Harmonic k -Nearest Neighbor Detector.	216
6.2.5	Local Outlier Factor (LOF).	217
6.3	A Review of Clustering, Histograms, and Density-Based Methods	219
6.3.1	Histogram and Clustering Methods	219
6.3.2	Kernel Density Methods	224
6.4	A Review of Dependency-Oriented Detectors.	225
6.4.1	Soft PCA: The Mahalanobis Method	226
6.4.2	Kernel Mahalanobis Method.	231
6.4.3	Decomposing Unsupervised Learning into Supervised Learning Problems	239
6.4.4	High-Dimensional Outliers Based on Group-Wise Dependencies	242
6.5	The Hidden Wildcard of Algorithm Parameters	243
6.5.1	Variable Subsampling and the Tyranny of Parameter Choice.	245
6.6	TRINITY: A Blend of Heterogeneous Base Detectors	247
6.7	Analysis of Performance.	248
6.7.1	Data Set Descriptions	249
6.7.2	Specific Details of Setting.	251
6.7.3	Summary of Findings	253
6.7.4	The Great Equalizing Power of Ensembles	260
6.7.5	The Argument for a Heterogeneous Combination	264
6.7.6	Discussion.	269
6.8	Conclusions	271
	References.	271
	Index	275

About the Authors



Charu C. Aggarwal is a Distinguished Research Staff Member (DRSM) at the IBM T. J. Watson Research Center in Yorktown Heights, New York. He completed his undergraduate degree in computer science from the Indian Institute of Technology at Kanpur in 1993 and his Ph.D. from the Massachusetts Institute of Technology in 1996.

He has worked extensively in the field of data mining. He has published more than 300 papers in refereed conferences and journals and authored over 80 patents. He is the author or editor of 16 books, including textbooks on data mining, recommender systems, and outlier analysis. Because of the commercial value of his patents, he has thrice been designated a Master Inventor at IBM. He is a recipient of an IBM Corporate Award (2003) for his work on bioterrorist threat detection in data streams, a recipient of the IBM Outstanding Innovation Award (2008) for his scientific contributions to privacy technology, a recipient of two IBM Outstanding Technical Achievement Awards (2009, 2015) for his work on data streams and high-dimensional data, respectively. He received the EDBT 2014 Test of Time Award for his work on condensation-based privacy-preserving data mining. He is also a recipient of the IEEE ICDM Research Contributions Award (2015), which is one of the two highest awards for influential research contributions in the field of data mining.

He has served as the general co-chair of the IEEE Big Data Conference (2014) and as the program

co-chair of the ACM CIKM Conference (2015), the IEEE ICDM Conference (2015), and the ACM KDD Conference (2016). He served as an associate editor of the IEEE Transactions on Knowledge and Data Engineering from 2004 to 2008. He is an associate editor of the ACM Transactions on Knowledge Discovery from Data, an associate editor of the IEEE Transactions on Big Data, an action editor of the Data Mining and Knowledge Discovery Journal, editor-in-chief of the ACM SIGKDD Explorations, and an associate editor of the Knowledge and Information Systems Journal. He serves on the advisory board of the Lecture Notes on Social Networks, a publication by Springer. He has served as the vice president of the SIAM Activity Group on data mining and is a member of the SIAM industry committee. He is a fellow of the SIAM, ACM, and the IEEE, for “contributions to knowledge discovery and data mining algorithms.”



Saket Sathé has worked at IBM Research Australia/USA since 2013. Saket received a Ph.D. degree in computer science from EPFL (Lausanne) in 2013. At EPFL, he was associated with the Distributed Information Systems Laboratory. Before that he received a Master’s (M.Tech.) degree in Electrical Engineering from the Indian Institute of Technology at Bombay in 2006. Prior to joining EPFL, he spent one year working for a start-up. His primary areas of interest are data mining and data management. Saket has served on program committees of several top-ranked conferences and has been invited to review papers for prominent peer-reviewed journals. His research has led to more than 20 papers and 5 patents. His work on sensor data management received the runner-up Best Paper Award in IEEE CollaborateCom, 2014. He is a member of the ACM, IEEE, and the SIAM.