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

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 Charu C. Aggarwal
January 2017 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 Sathe 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 Sathe 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

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 Sathe 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.