Outlier Analysis

Charu C. Aggarwal

Outlier Analysis

Second Edition

Springer

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

ISBN 978-3-319-47577-6 ISBN 978-3-319-47578-3 (eBook)
DOI 10.1007/978-3-319-47578-3

Library of Congress Control Number: 2016961247

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

To my wife, my daughter Sayani,
and my late parents Dr. Prem Sarup and Mrs. Pushplata Aggarwal.

Contents

Preface

"All things excellent are as difficult as they are rare."– Baruch Spinoza

First Edition

Most of the earliest work on outlier detection was performed by the statistics community. While statistical methods are mathematically more precise, they have several shortcomings, such as simplified assumptions about data representations, poor algorithmic scalability, and a low focus on interpretability. With the increasing advances in hardware technology for *data collection*, and advances in software technology (databases) for *data organization*, computer scientists have increasingly been participating in the latest advancements of this field. Computer scientists approach this field based on their practical experiences in managing large amounts of data, and with far fewer assumptions– the data can be of any type, structured or unstructured, and may be extremely large. Furthermore, issues such as computational efficiency and intuitive analysis of the data are generally considered more important by computer scientists than mathematical precision, though the latter is important as well. This is the approach of professionals from the field of data mining, an area of computer science that was founded about 20 years ago. This has led to the formation of multiple academic communities on the subject, which have remained separated, partially because of differences in technical style and opinions about the importance of different problems and approaches to the subject. At this point, data mining professionals (with a computer science background) are much more actively involved in this area as compared to statisticians. This seems to be a major change in the research landscape. This book presents outlier detection from an integrated perspective, though the focus is towards computer science professionals. Special emphasis was placed on relating the methods from different communities with one another.

The key advantage of writing the book at this point in time is that the vast amount of work done by computer science professionals in the last two decades has remained largely untouched by a formal book on the subject. The classical books relevant to outlier analysis are as follows:

- P. Rousseeuw and A. Leroy. Robust Regression and Outlier Detection, *Wiley*, 2003.
- V. Barnett and T. Lewis. Outliers in Statistical Data, *Wiley*, 1994.
- D. Hawkins. Identification of Outliers, *Chapman and Hall*, 1980.

We note that these books are quite outdated, and the most recent among them is a decade old. Furthermore, this (most recent) book is really focused on the relationship between regression and outlier analysis, rather than the latter. Outlier analysis is a much broader area, in which regression analysis is only a small part. The other books are even older, and are between 15 and 25 years old. They are exclusively targeted to the statistics community. This is not surprising, given that the first mainstream computer science conference in data mining (KDD) was organized in 1995. Most of the work in the data-mining community was performed after the writing of these books. Therefore, many key topics of interest to the broader data mining community are not covered in these books. Given that outlier analysis has been explored by a much broader community, including databases, data mining, statistics, and machine learning, we feel that our book incorporates perspectives from a much broader audience and brings together different points of view.

The chapters of this book have been organized carefully, with a view of covering the area extensively in a natural order. Emphasis was placed on simplifying the content, so that students and practitioners can also benefit from the book. While we did not originally intend to create a textbook on the subject, it evolved during the writing process into a work that can also be used as a teaching aid. Furthermore, it can also be used as a reference book, since each chapter contains extensive bibliographic notes. Therefore, this book serves a dual purpose, providing a comprehensive exposition of the topic of outlier detection from multiple points of view.

Additional Notes for the Second Edition

The second edition of this book is a significant enhancement over the first edition. In particular, most of the chapters have been upgraded with new material and recent techniques. More explanations have been added at several places and newer techniques have also been added. An entire chapter on outlier ensembles has been added. Many new topics have been added to the book such as feature selection, one-class support vector machines, one-class neural networks, matrix factorization, spectral methods, wavelet transforms, and supervised learning. Every chapter has been updated with the latest algorithms on the topic.

Last but not least, the first edition was classified by the publisher as a monograph, whereas the second edition is formally classified as a textbook. The writing style has been enhanced to be easily understandable to students. Many algorithms have been described in greater detail, as one might expect from a textbook. It is also accompanied with a solution manual for classroom teaching.

Acknowledgments

First Edition

I would like to thank my wife and daughter for their love and support during the writing of this book. The writing of a book requires significant time that is taken away from family members. This book is the result of their patience with me during this time. I also owe my late parents a debt of gratitude for instilling in me a love of education, which has played an important inspirational role in my book-writing efforts.

I would also like to thank my manager Nagui Halim for providing the tremendous support necessary for the writing of this book. His professional support has been instrumental for my many book efforts in the past and present.

Over the years, I have benefited from the insights of numerous collaborators. An incomplete list of these long-term collaborators in alphabetical order is Tarek F. Abdelzaher, Jiawei Han, Thomas S. Huang, Latifur Khan, Mohammad M. Masud, Spiros Papadimitriou, Guojun Qi, and Philip S. Yu. I would like to thank them for their collaborations and insights over the course of many years.

I would also like to specially thank my advisor James B. Orlin for his guidance during my early years as a researcher. While I no longer work in the same area, the legacy of what I learned from him is a crucial part of my approach to research. In particular, he taught me the importance of intuition and simplicity of thought in the research process. These are more important aspects of research than is generally recognized. This book is written in a simple and intuitive style, and is meant to improve accessibility of this area to both researchers and practitioners.

Finally, I would like to thank Lata Aggarwal for helping me with some of the figures created using PowerPoint graphics in this book.

Acknowledgments for Second Edition

I received significant feedback from various colleagues during the writing of the second edition. In particular, I would like to acknowledge Leman Akoglu, Chih-Jen Lin, Saket Sathe, Jiliang Tang, and Suhang Wang. Leman and Saket provided detailed feedback on several sections and chapters of this book.

Author Biography

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 15 books, including a textbook on data mining and a comprehensive book on 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 bio-terrorist 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."

Chapter 1

An Introduction to Outlier Analysis

> "Never take the comment that you are different as a condemnation, it might be a compliment. It might mean that you possess unique qualities that, like the most rarest of diamonds is ... one of a kind." – Eugene Nathaniel Butler

1.1 Introduction

An outlier is a data point that is significantly different from the remaining data. Hawkins defined [249] an outlier as follows:

> "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism."

Outliers are also referred to as *abnormalities*, *discordants*, *deviants*, or *anomalies* in the data mining and statistics literature. In most applications, the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the generating process behaves unusually, it results in the creation of outliers. Therefore, an outlier often contains useful information about abnormal characteristics of the systems and entities that impact the data generation process. The recognition of such unusual characteristics provides useful application-specific insights. Some examples are as follows:

- **Intrusion detection systems:** In many computer systems, different types of data are collected about the operating system calls, network traffic, or other user actions. This data may show unusual behavior because of malicious activity. The recognition of such activity is referred to as intrusion detection.
- **Credit-card fraud:** Credit-card fraud has become increasingly prevalent because of greater ease with which sensitive information such as a credit-card number can be compromised. In many cases, unauthorized use of a credit card may show different patterns, such as buying sprees from particular locations or very large transactions. Such patterns can be used to detect outliers in credit-card transaction data.

C.C. Aggarwal, *Outlier Analysis*, DOI 10.1007/978-3-319-47578-3_1

- **Interesting sensor events:** Sensors are often used to track various environmental and location parameters in many real-world applications. Sudden changes in the underlying patterns may represent events of interest. Event detection is one of the primary motivating applications in the field of sensor networks. As discussed later in this book, event detection is an important *temporal* version of outlier detection.

- **Medical diagnosis:** In many medical applications, the data is collected from a variety of devices such as magnetic resonance imaging (MRI) scans, positron emission tomography (PET) scans or electrocardiogram (ECG) time-series. Unusual patterns in such data typically reflect disease conditions.

- **Law enforcement:** Outlier detection finds numerous applications in law enforcement, especially in cases where unusual patterns can only be discovered over time through multiple actions of an entity. Determining fraud in financial transactions, trading activity, or insurance claims typically requires the identification of unusual patterns in the data generated by the actions of the criminal entity.

- **Earth science:** A significant amount of spatiotemporal data about weather patterns, climate changes, or land-cover patterns is collected through a variety of mechanisms such as satellites or remote sensing. Anomalies in such data provide significant insights about human activities or environmental trends that may be the underlying causes.

In all these applications, the data has a "normal" model, and anomalies are recognized as deviations from this normal model. Normal data points are sometimes also referred to as *inliers*. In some applications such as intrusion or fraud detection, outliers correspond to *sequences* of multiple data points rather than individual data points. For example, a fraud event may often reflect the actions of an individual in a particular sequence. The specificity of the sequence is relevant to identifying the anomalous event. Such anomalies are also referred to as *collective anomalies*, because they can only be inferred collectively from a set or sequence of data points. Such collective anomalies are often a result of unusual *events* that generate anomalous patterns of activity. This book will address these different types of anomalies.

The output of an outlier detection algorithm can be one of two types:

- **Outlier scores:** Most outlier detection algorithms output a score quantifying the level of "outlierness" of each data point. This score can also be used to rank the data points in order of their outlier tendency. This is a very general form of output, which retains all the information provided by a particular algorithm, but it does not provide a concise summary of the small number of data points that should be considered outliers.

- **Binary labels:** A second type of output is a binary label indicating whether a data point is an outlier or not. Although some algorithms might directly return binary labels, outlier scores can also be converted into binary labels. This is typically achieved by imposing thresholds on outlier scores, and the threshold is chosen based on the statistical distribution of the scores. A binary labeling contains less information than a scoring mechanism, but it is the final result that is often needed for decision making in practical applications.

It is often a subjective judgement, as to what constitutes a "sufficient" deviation for a point to be considered an outlier. In real applications, the data may be embedded in a