
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We note that these books are quite outdated, and the most recent among them is a decade old.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.

Outlier Analysis

Aggarwal, C.C.

2017, XXII, 466 p. 78 illus., 13 illus. in color., Hardcover

ISBN: 978-3-319-47577-6