Preface

Streaming data analysis in real time is becoming the standard to obtain useful knowledge from what is happening right now, allowing organizations to react quickly when problems appear, or to detect new trends, helping them to improve their performance. This book presents many of the algorithms and techniques that are currently used in the field of data stream mining. A software framework that implements many of the techniques explained in this book is available from the Web as the open-source project called MOA.

The goal of this book is to present the techniques in data stream mining to three specific groups of readers:

1. Readers who want to use stream mining as a tool, who do not have a strong background in algorithmics or programming, but do have a basic background in data mining. An example would be students or professionals in fields such as management, business intelligence, or marketing. We provide a hands-on introduction to MOA, in a task-oriented (not algorithm-oriented) way.
2. Readers who want to do research or innovation in data stream mining. They would like to know details of the algorithms, evaluation methods, and so on, in order to create new algorithms or use existing ones, evaluate their performance, and possibly include them in their applications. This group comprises advanced undergraduate, master's, and PhD students in computing or data science, as well as developers in innovative environments.
3. Readers who, in addition to the above, want to try including new algorithms in MOA, possibly contributing them to the project. They need to know the class structure of MOA and how to create, for instance, new learners for MOA.

To achieve this goal, the book is divided in three parts. Part I is a quick introduction to big data stream mining. It is structured in three chapters: two that introduce big data mining and basic methodologies for mining data streams, and a hands-on chapter on using MOA for readers who prefer to get started and explore on their own.

For a longer course on data stream mining, part II of the book presents a detailed explanation of the problems in data stream mining and the most important algorithms. Since this is a vast area, some priority has been given to the methods that have been implemented in MOA. It starts with a chapter covering sketching techniques, which in our opinion deserve to be better known (and

used) by the stream mining community. Most of the chapters contain a set of exercises or an MOA-based lab session, or both.

Finally, part III is devoted to the MOA software. It covers its use via the graphical user interface and via the command line, and moves to using MOA via its API, and implementing new methods within MOA.

Readers of type 1 should read part I, possibly chapter 11 for a broad view of MOA's ecosystem, and then chapter 12 for other options available from the MOA GUI.

Readers of type 2 should read part I, at least sections 4.1 to 4.3 (and more of chapter 4 if they are interested in sketches), chapter 5, and chapter 6. After that, they can read chapters 7 to 10 pretty much independently according to their interests. Then they should continue to chapters 11 to 14 if they plan to call MOA from their applications.

Readers of type 3 should in addition read Chapter 15.

The accompanying website

`https://mitpress.mit.edu/books/data-stream-mining`

will contain updates and corrections to the book, slides, additional sets of exercises and lab sessions, and other course material. Contributions by readers are welcome.

Several books on data stream mining have emerged over the last decade. The books edited by Garofalakis, Gehrke, and Rastogi on data stream management [118], and by Aggarwal on data streams [4], cover some common topics with the material presented here, but the perspective of these books is more from the very-large-database community rather than from the data mining or machine learning communities.

The latter perspective is very much present in the book by Gama [110], who covers a similar territory but does not include a common framework for development and evaluation as provided by MOA. Rather, the book presents pseudo-code of algorithms, some of which are implemented in MOA and some not. As such, it is a very useful companion to this book.

To keep up with this rapidly developing field, we recommend regular reading of the proceedings of the following conferences: Knowledge Discovery in Databases (KDD), International Conference on Data Mining (ICDM), Symposium on Applied Computing (SAC) – has a track on data streams, European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), SIAM Conference on Data Mining (SDM), and Data Science and Advanced Analytics (DSAA).

To date, there is no dedicated journal on data stream mining, so articles appear on the topic across a number of journals too numerous to list.

Acknowledgments. We would like to thank the following groups of people, who have contributed to this book and the software behind it. It is not possible to mention all by name, either because the names are unknown to us (e.g., the reviewers) or because the list is long and we may inadvertently miss someone (e.g., our coauthors or the many students and other people who have contributed to MOA by asking questions, pointing out bugs, and so forth, on the mailing list, or those who have directly contributed code).

We would like to thank the people at MIT Press, and in particular Marie Lufkin Lee, Christine Bridget Savage, and Kathleen Hensley, for their assistance.

It is worth acknowledging that the inspiration for this project and book came from the groundbreaking work of the WEKA project.

For those authors working in the area of stream mining, we would like to apologize in advance if your work is not mentioned in the book. Such a state of affairs will have arisen because of space limitations, ignorance, or the wrong choice on our part.

Work by Ricard Gavaldà has been partially supported by the MACDA project of Generalitat de Catalunya (SGR2014-0890) and by the APCOM project of MINECO (TIN2014-57226).