
MACHINE LEARNING FOR DATA STREAMS

with Practical Examples in MOA

Adaptive Computation and Machine Learning

Francis Bach, Editor

Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns,
Associate Editors

A complete list of books published in The Adaptive Computation and Machine Learning series appears at the back of this book.

MACHINE LEARNING FOR DATA STREAMS
with Practical Examples in MOA

Albert Bifet
Ricard Gavaldà
Geoff Holmes
Bernhard Pfahringer

The MIT Press
Cambridge, Massachusetts
London, England

© 2017 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Times Roman and Mathtime Pro 2 by the authors.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data is available

ISBN: 978-0-262-03779-2

10 9 8 7 6 5 4 3 2 1

Contents

List of Figures	xiii	
List of Tables	xvii	
Preface	xix	
I	INTRODUCTION	1
1	Introduction	3
1.1	Big Data	3
1.1.1	Tools: Open-Source Revolution	5
1.1.2	Challenges in Big Data	6
1.2	Real-Time Analytics	8
1.2.1	Data Streams	8
1.2.2	Time and Memory	8
1.2.3	Applications	8
1.3	What This Book Is About	10
2	Big Data Stream Mining	11
2.1	Algorithms	11
2.2	Classification	12
2.2.1	Classifier Evaluation in Data Streams	14
2.2.2	Majority Class Classifier	15
2.2.3	No-Change Classifier	15
2.2.4	Lazy Classifier	15
2.2.5	Naive Bayes	16
2.2.6	Decision Trees	16
2.2.7	Ensembles	17
2.3	Regression	17
2.4	Clustering	17
2.5	Frequent Pattern Mining	18
3	Hands-on Introduction to MOA	21
3.1	Getting Started	21
3.2	The Graphical User Interface for Classification	23
3.2.1	Drift Stream Generators	25
3.3	Using the Command Line	29

II	STREAM MINING	33
4	Streams and Sketches	35
4.1	Setting: Approximation Algorithms	35
4.2	Concentration Inequalities	37
4.3	Sampling	39
4.4	Counting Total Items	41
4.5	Counting Distinct Elements	42
4.5.1	Linear Counting	43
4.5.2	Cohen's Logarithmic Counter	44
4.5.3	The Flajolet-Martin Counter and HyperLogLog	45
4.5.4	An Application: Computing Distance Functions in Graphs	47
4.5.5	Discussion: Log vs. Linear	48
4.6	Frequency Problems	48
4.6.1	The SPACESAVING Sketch	49
4.6.2	The CM-Sketch Algorithm	51
4.6.3	CountSketch	54
4.6.4	Moment Computation	56
4.7	Exponential Histograms for Sliding Windows	57
4.8	Distributed Sketching: Mergeability	60
4.9	Some Technical Discussions and Additional Material	61
4.9.1	Hash Functions	61
4.9.2	Creating (ϵ, δ) -Approximation Algorithms	62
4.9.3	Other Sketching Techniques	63
4.10	Exercises	63
5	Dealing with Change	67
5.1	Notion of Change in Streams	67
5.2	Estimators	72
5.2.1	Sliding Windows and Linear Estimators	73
5.2.2	Exponentially Weighted Moving Average	73
5.2.3	Unidimensional Kalman Filter	74
5.3	Change Detection	75
5.3.1	Evaluating Change Detection	75
5.3.2	The CUSUM and Page-Hinkley Tests	75

5.3.3	Statistical Tests	76
5.3.4	Drift Detection Method	78
5.3.5	ADWIN	79
5.4	Combination with Other Sketches and Multidimensional Data	81
5.5	Exercises	81
6	Classification	85
6.1	Classifier Evaluation	86
6.1.1	Error Estimation	87
6.1.2	Distributed Evaluation	88
6.1.3	Performance Evaluation Measures	90
6.1.4	Statistical Significance	92
6.1.5	A Cost Measure for the Mining Process	93
6.2	Baseline Classifiers	94
6.2.1	Majority Class	94
6.2.2	No-change Classifier	94
6.2.3	Naive Bayes	95
6.2.4	Multinomial Naive Bayes	98
6.3	Decision Trees	99
6.3.1	Estimating Split Criteria	101
6.3.2	The Hoeffding Tree	102
6.3.3	CVFDT	105
6.3.4	VFDTc and UFFT	107
6.3.5	Hoeffding Adaptive Tree	108
6.4	Handling Numeric Attributes	109
6.4.1	VFML	110
6.4.2	Exhaustive Binary Tree	110
6.4.3	Greenwald and Khanna's Quantile Summaries	111
6.4.4	Gaussian Approximation	111
6.5	Perceptron	113
6.6	Lazy Learning	114
6.7	Multi-label Classification	115
6.7.1	Multi-label Hoeffding Trees	116

6.8	Active Learning	117
6.8.1	Random Strategy	119
6.8.2	Fixed Uncertainty Strategy	119
6.8.3	Variable Uncertainty Strategy	119
6.8.4	Uncertainty Strategy with Randomization	121
6.9	Concept Evolution	121
6.10	Lab Session with MOA	122
7	Ensemble Methods	129
7.1	Accuracy-Weighted Ensembles	129
7.2	Weighted Majority	130
7.3	Stacking	132
7.4	Bagging	133
7.4.1	Online Bagging Algorithm	133
7.4.2	Bagging with a Change Detector	133
7.4.3	Leveraging Bagging	134
7.5	Boosting	135
7.6	Ensembles of Hoeffding Trees	136
7.6.1	Hoeffding Option Trees	136
7.6.2	Random Forests	136
7.6.3	Perceptron Stacking of Restricted Hoeffding Trees	137
7.6.4	Adaptive-Size Hoeffding Trees	138
7.7	Recurrent Concepts	139
7.8	Lab Session with MOA	139
8	Regression	143
8.1	Introduction	143
8.2	Evaluation	144
8.3	Perceptron Learning	145
8.4	Lazy Learning	145
8.5	Decision Tree Learning	146
8.6	Decision Rules	146
8.7	Regression in MOA	148

9	Clustering	149
9.1	Evaluation Measures	150
9.2	The k -means Algorithm	151
9.3	BIRCH, BICO, and CLUSTREAM	152
9.4	Density-Based Methods: DBSCAN and Den-Stream	154
9.5	CLUSTREE	156
9.6	StreamKM++: Coresets	158
9.7	Additional Material	159
9.8	Lab Session with MOA	160
10	Frequent Pattern Mining	165
10.1	An Introduction to Pattern Mining	165
10.1.1	Patterns: Definitions and Examples	165
10.1.2	Batch Algorithms for Frequent Pattern Mining	168
10.1.3	Closed and Maximal Patterns	169
10.2	Frequent Pattern Mining in Streams: Approaches	170
10.2.1	Coresets of Closed Patterns	172
10.3	Frequent Itemset Mining on Streams	174
10.3.1	Reduction to Heavy Hitters	174
10.3.2	Moment	174
10.3.3	FP-STREAM	175
10.3.4	IncMine	176
10.4	Frequent Subgraph Mining on Streams	178
10.4.1	WINGRAPHMINER	179
10.4.2	ADAGRAPHMINER	179
10.5	Additional Material	181
10.6	Exercises	182
III	THE MOA SOFTWARE	185
11	Introduction to MOA and Its Ecosystem	187
11.1	MOA Architecture	188
11.2	Installation	188
11.3	Recent Developments in MOA	188
11.4	Extensions to MOA	189

11.5	ADAMS	190
11.6	MEKA	193
11.7	OpenML	194
11.8	StreamDM	195
11.9	Streams	196
11.10	Apache SAMOA	196
12	The Graphical User Interface	201
12.1	Getting Started with the GUI	201
12.2	Classification and Regression	201
12.2.1	Tasks	203
12.2.2	Data Feeds and Data Generators	204
12.2.3	Bayesian Classifiers	208
12.2.4	Decision Trees	208
12.2.5	Meta Classifiers (Ensembles)	209
12.2.6	Function Classifiers	210
12.2.7	Drift Classifiers	210
12.2.8	Active Learning Classifiers	211
12.3	Clustering	211
12.3.1	Data Feeds and Data Generators	212
12.3.2	Stream Clustering Algorithms	212
12.3.3	Visualization and Analysis	212
13	Using the Command Line	217
13.1	Learning Task for Classification and Regression	217
13.2	Evaluation Tasks for Classification and Regression	217
13.3	Learning and Evaluation Tasks for Classification and Regression	218
13.4	Comparing Two Classifiers	219
14	Using the API	221
14.1	MOA Objects	221
14.2	Options	221
14.3	Prequential Evaluation Example	224

15	Developing New Methods in MOA	227
	15.1 Main Classes in MOA	227
	15.2 Creating a New Classifier	228
	15.3 Compiling a Classifier	237
	15.4 Good Programming Practices in MOA	237
	Bibliography	239
	Index	257

This is a section of [doi:10.7551/mitpress/10654.001.0001](https://doi.org/10.7551/mitpress/10654.001.0001)

Machine Learning for Data Streams

with Practical Examples in MOA

By: Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, Bernhard Pfahringer

Citation:

Machine Learning for Data Streams: with Practical Examples in MOA

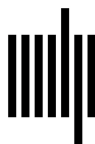
By: Albert Bifet, Ricard Gavaldà, Geoffrey Holmes, Bernhard Pfahringer

DOI: 10.7551/mitpress/10654.001.0001

ISBN (electronic): 9780262346047

Publisher: The MIT Press

Published: 2023



The MIT Press

© 2017 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form or by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

This book was set in Times Roman and Mathtime Pro 2 by the authors.

Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data is available

ISBN: 978-0-262-03779-2

10 9 8 7 6 5 4 3 2 1