
DATA STREAMS:
MODELS AND ALGORITHMS

DATA STREAMS: MODELS AND ALGORITHMS

Edited by
CHARU C. AGGARWAL
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598

Kluwer Academic Publishers
Boston/Dordrecht/London

Contents

List of Figures	xi
List of Tables	xv
Preface	xvii
1	
An Introduction to Data Streams	1
<i>Charu C. Aggarwal</i>	
1. Introduction	1
2. Stream Mining Algorithms	2
3. Conclusions and Summary	6
References	7
2	
On Clustering Massive Data Streams: A Summarization Paradigm	9
<i>Charu C. Aggarwal, Jiawei Han, Jianyong Wang and Philip S. Yu</i>	
1. Introduction	10
2. The Micro-clustering Based Stream Mining Framework	12
3. Clustering Evolving Data Streams: A Micro-clustering Approach	17
3.1 Micro-clustering Challenges	18
3.2 Online Micro-cluster Maintenance: The CluStream Algorithm	19
3.3 High Dimensional Projected Stream Clustering	22
4. Classification of Data Streams: A Micro-clustering Approach	23
4.1 On-Demand Stream Classification	24
5. Other Applications of Micro-clustering and Research Directions	26
6. Performance Study and Experimental Results	27
7. Discussion	36
References	36
3	
A Survey of Classification Methods in Data Streams	39
<i>Mohamed Medhat Gaber, Arkady Zaslavsky and Shonali Krishnaswamy</i>	
1. Introduction	39
2. Research Issues	41
3. Solution Approaches	43
4. Classification Techniques	44
4.1 Ensemble Based Classification	45
4.2 Very Fast Decision Trees (VFDT)	46

4.3	On Demand Classification	48
4.4	Online Information Network (OLIN)	48
4.5	LWClass Algorithm	49
4.6	ANNCAD Algorithm	51
4.7	SCALLOP Algorithm	51
5.	Summary	52
	References	53
4	Frequent Pattern Mining in Data Streams	61
	<i>Ruoming Jin and Gagan Agrawal</i>	
1.	Introduction	61
2.	Overview	62
3.	New Algorithm	67
4.	Work on Other Related Problems	79
5.	Conclusions and Future Directions	80
	References	81
5	A Survey of Change Diagnosis Algorithms in Evolving Data Streams	85
	<i>Charu C. Aggarwal</i>	
1.	Introduction	86
2.	The Velocity Density Method	88
	2.1 Spatial Velocity Profiles	93
	2.2 Evolution Computations in High Dimensional Case	95
	2.3 On the use of clustering for characterizing stream evolution	96
3.	On the Effect of Evolution in Data Mining Algorithms	97
4.	Conclusions	100
	References	101
6	Multi-Dimensional Analysis of Data Streams Using Stream Cubes	103
	<i>Jiawei Han, Y. Dora Cai, Yixin Chen, Guozhu Dong, Jian Pei, Benjamin W. Wah, and Jianyong Wang</i>	
1.	Introduction	104
2.	Problem Definition	106
3.	Architecture for On-line Analysis of Data Streams	108
	3.1 Tilted time frame	108
	3.2 Critical layers	110
	3.3 Partial materialization of stream cube	111
4.	Stream Data Cube Computation	112
	4.1 Algorithms for cube computation	115
5.	Performance Study	117
6.	Related Work	120
7.	Possible Extensions	121
8.	Conclusions	122
	References	123

<i>Contents</i>	vii
7	
Load Shedding in Data Stream Systems	127
<i>Brian Babcock, Mayur Datar and Rajeev Motwani</i>	
1. Load Shedding for Aggregation Queries	128
1.1 Problem Formulation	129
1.2 Load Shedding Algorithm	133
1.3 Extensions	141
2. Load Shedding in Aurora	142
3. Load Shedding for Sliding Window Joins	144
4. Load Shedding for Classification Queries	145
5. Summary	146
References	146
8	
The Sliding-Window Computation Model and Results	149
<i>Mayur Datar and Rajeev Motwani</i>	
0.1 Motivation and Road Map	150
1. A Solution to the BASICCOUNTING Problem	152
1.1 The Approximation Scheme	154
2. Space Lower Bound for BASICCOUNTING Problem	157
3. Beyond 0's and 1's	158
4. References and Related Work	163
5. Conclusion	164
References	166
9	
A Survey of Synopsis Construction in Data Streams	169
<i>Charu C. Aggarwal, Philip S. Yu</i>	
1. Introduction	169
2. Sampling Methods	172
2.1 Random Sampling with a Reservoir	174
2.2 Concise Sampling	176
3. Wavelets	177
3.1 Recent Research on Wavelet Decomposition in Data Streams	182
4. Sketches	184
4.1 Fixed Window Sketches for Massive Time Series	185
4.2 Variable Window Sketches of Massive Time Series	185
4.3 Sketches and their applications in Data Streams	186
4.4 Sketches with p -stable distributions	190
4.5 The Count-Min Sketch	191
4.6 Related Counting Methods: Hash Functions for Determining Distinct Elements	193
4.7 Advantages and Limitations of Sketch Based Methods	194
5. Histograms	196
5.1 One Pass Construction of Equi-depth Histograms	198
5.2 Constructing V-Optimal Histograms	198
5.3 Wavelet Based Histograms for Query Answering	199
5.4 Sketch Based Methods for Multi-dimensional Histograms	200
6. Discussion and Challenges	200

References	202
10	
A Survey of Join Processing in Data Streams	209
<i>Junyi Xie and Jun Yang</i>	
1. Introduction	209
2. Model and Semantics	210
3. State Management for Stream Joins	213
3.1 Exploiting Constraints	214
3.2 Exploiting Statistical Properties	216
4. Fundamental Algorithms for Stream Join Processing	225
5. Optimizing Stream Joins	227
6. Conclusion	230
Acknowledgments	232
References	232
11	
Indexing and Querying Data Streams	237
<i>Ahmet Bulut, Ambuj K. Singh</i>	
1. Introduction	238
2. Indexing Streams	239
2.1 Preliminaries and definitions	239
2.2 Feature extraction	240
2.3 Index maintenance	244
2.4 Discrete Wavelet Transform	246
3. Querying Streams	248
3.1 Monitoring an aggregate query	248
3.2 Monitoring a pattern query	251
3.3 Monitoring a correlation query	252
4. Related Work	254
5. Future Directions	255
5.1 Distributed monitoring systems	255
5.2 Probabilistic modeling of sensor networks	256
5.3 Content distribution networks	256
6. Chapter Summary	257
References	257
12	
Dimensionality Reduction and Forecasting on Streams	261
<i>Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos</i>	
1. Related work	264
2. Principal component analysis (PCA)	265
3. Auto-regressive models and recursive least squares	267
4. MUSCLES	269
5. Tracking correlations and hidden variables: SPIRIT	271
6. Putting SPIRIT to work	276
7. Experimental case studies	278

<i>Contents</i>	ix
8. Performance and accuracy	283
9. Conclusion	286
Acknowledgments	286
References	287
13	
A Survey of Distributed Mining of Data Streams	289
<i>Srinivasan Parthasarathy, Amol Ghoting and Matthew Eric Otey</i>	
1. Introduction	289
2. Outlier and Anomaly Detection	291
3. Clustering	295
4. Frequent itemset mining	296
5. Classification	297
6. Summarization	298
7. Mining Distributed Data Streams in Resource Constrained Environ- ments	299
8. Systems Support	300
References	304
14	
Algorithms for Distributed Data Stream Mining	309
<i>Kanishka Bhaduri, Kamalika Das, Krishnamoorthy Sivakumar, Hillol Kargupta, Ran Wolff and Rong Chen</i>	
1. Introduction	310
2. Motivation: Why Distributed Data Stream Mining?	311
3. Existing Distributed Data Stream Mining Algorithms	312
4. A <i>local</i> algorithm for distributed data stream mining	315
4.1 Local Algorithms : definition	315
4.2 Algorithm details	316
4.3 Experimental results	318
4.4 Modifications and extensions	320
5. Bayesian Network Learning from Distributed Data Streams	321
5.1 Distributed Bayesian Network Learning Algorithm	322
5.2 Selection of samples for transmission to global site	323
5.3 Online Distributed Bayesian Network Learning	324
5.4 Experimental Results	326
6. Conclusion	326
References	329
15	
A Survey of Stream Processing Problems and Techniques in Sensor Networks	333
<i>Sharmila Subramaniam, Dimitrios Gunopulos</i>	
1. Challenges	334

2.	The Data Collection Model	335
3.	Data Communication	335
4.	Query Processing	337
	4.1 Aggregate Queries	338
	4.2 Join Queries	340
	4.3 Top- k Monitoring	341
	4.4 Continuous Queries	341
5.	Compression and Modeling	342
	5.1 Data Distribution Modeling	343
	5.2 Outlier Detection	344
6.	Application: Tracking of Objects using Sensor Networks	345
7.	Summary	347
	References	348
	Index	353