
Contents

List of Tables	xi
List of Figures	xiii
List of Algorithms	xv
Foreword	xvii
Acknowledgments	xix
1 Knowledge Discovery from Data Streams	1
1.1 Introduction	1
1.2 An Illustrative Example	2
1.3 A World in Movement	4
1.4 Data Mining and Data Streams	5
2 Introduction to Data Streams	7
2.1 Data Stream Models	7
2.1.1 Research Issues in Data Stream Management Systems	8
2.1.2 An Illustrative Problem	8
2.2 Basic Streaming Methods	9
2.2.1 Illustrative Examples	10
2.2.1.1 Counting the Number of Occurrences of the Elements in a Stream	10
2.2.1.2 Counting the Number of Distinct Values in a Stream	11
2.2.2 Bounds of Random Variables	11
2.2.3 Poisson Processes	13
2.2.4 Maintaining Simple Statistics from Data Streams . . .	14
2.2.5 Sliding Windows	14
2.2.5.1 Computing Statistics over Sliding Windows: The ADWIN Algorithm	16
2.2.6 Data Synopsis	19
2.2.6.1 Sampling	19
2.2.6.2 Synopsis and Histograms	20
2.2.6.3 Wavelets	21
2.2.6.4 Discrete Fourier Transform	22

2.3	Illustrative Applications	23
2.3.1	A Data Warehouse Problem: Hot-Lists	23
2.3.2	Computing the Entropy in a Stream	24
2.3.3	Monitoring Correlations Between Data Streams	27
2.3.4	Monitoring Threshold Functions over Distributed Data Streams	29
2.4	Notes	30
3	Change Detection	33
3.1	Introduction	33
3.2	Tracking Drifting Concepts	34
3.2.1	The Nature of Change	35
3.2.2	Characterization of Drift Detection Methods	36
3.2.2.1	Data Management	37
3.2.2.2	Detection Methods	38
3.2.2.3	Adaptation Methods	40
3.2.2.4	Decision Model Management	41
3.2.3	A Note on Evaluating Change Detection Methods	41
3.3	Monitoring the Learning Process	42
3.3.1	Drift Detection Using Statistical Process Control	42
3.3.2	An Illustrative Example	45
3.4	Final Remarks	46
3.5	Notes	47
4	Maintaining Histograms from Data Streams	49
4.1	Introduction	49
4.2	Histograms from Data Streams	50
4.2.1	K-buckets Histograms	50
4.2.2	Exponential Histograms	51
4.2.2.1	An Illustrative Example	52
4.2.2.2	Discussion	52
4.3	The Partition Incremental Discretization Algorithm - PiD	53
4.3.1	Analysis of the Algorithm	56
4.3.2	Change Detection in Histograms	56
4.3.3	An Illustrative Example	57
4.4	Applications to Data Mining	59
4.4.1	Applying PiD in Supervised Learning	59
4.4.2	Time-Changing Environments	61
4.5	Notes	62
5	Evaluating Streaming Algorithms	63
5.1	Introduction	63
5.2	Learning from Data Streams	64
5.3	Evaluation Issues	65
5.3.1	Design of Evaluation Experiments	66

5.3.2	Evaluation Metrics	67
5.3.2.1	Error Estimators Using a Single Algorithm and a Single Dataset	68
5.3.2.2	An Illustrative Example	68
5.3.3	Comparative Assessment	69
5.3.3.1	The 0 – 1 Loss Function	70
5.3.3.2	Illustrative Example	71
5.3.4	Evaluation Methodology in Non-Stationary Environments	72
5.3.4.1	The Page-Hinkley Algorithm	72
5.3.4.2	Illustrative Example	73
5.4	Lessons Learned and Open Issues	75
5.5	Notes	77
6	Clustering from Data Streams	79
6.1	Introduction	79
6.2	Clustering Examples	80
6.2.1	Basic Concepts	80
6.2.2	Partitioning Clustering	82
6.2.2.1	The Leader Algorithm	82
6.2.2.2	Single Pass <i>k</i> -Means	82
6.2.3	Hierarchical Clustering	83
6.2.4	Micro Clustering	85
6.2.4.1	Discussion	86
6.2.4.2	Monitoring Cluster Evolution	86
6.2.5	Grid Clustering	87
6.2.5.1	Computing the Fractal Dimension	88
6.2.5.2	Fractal Clustering	88
6.3	Clustering Variables	90
6.3.1	A Hierarchical Approach	91
6.3.1.1	Growing the Hierarchy	91
6.3.1.2	Aggregating at Concept Drift Detection	94
6.3.1.3	Analysis of the Algorithm	96
6.4	Notes	96
7	Frequent Pattern Mining	97
7.1	Introduction to Frequent Itemset Mining	97
7.1.1	The Search Space	98
7.1.2	The FP-growth Algorithm	100
7.1.3	Summarizing Itemsets	100
7.2	Heavy Hitters	101
7.3	Mining Frequent Itemsets from Data Streams	103
7.3.1	Landmark Windows	104
7.3.1.1	The LossyCounting Algorithm	104
7.3.1.2	Frequent Itemsets Using LossyCounting	104

7.3.2	Mining Recent Frequent Itemsets	105
7.3.2.1	Maintaining Frequent Itemsets in Sliding Windows	105
7.3.2.2	Mining Closed Frequent Itemsets over Sliding Windows	106
7.3.3	Frequent Itemsets at Multiple Time Granularities	108
7.4	Sequence Pattern Mining	110
7.4.1	Reservoir Sampling for Sequential Pattern Mining over Data Streams	111
7.5	Notes	113
8	Decision Trees from Data Streams	115
8.1	Introduction	115
8.2	The Very Fast Decision Tree Algorithm	116
8.2.1	VFDT —The Base Algorithm	116
8.2.2	Analysis of the VFDT Algorithm	118
8.3	Extensions to the Basic Algorithm	119
8.3.1	Processing Continuous Attributes	119
8.3.1.1	Exhaustive Search	119
8.3.1.2	Discriminant Analysis	121
8.3.2	Functional Tree Leaves	123
8.3.3	Concept Drift	124
8.3.3.1	Detecting Changes	126
8.3.3.2	Reacting to Changes	127
8.3.4	Final Comments	128
8.4	OLIN: Info-Fuzzy Algorithms	129
8.5	Notes	132
9	Novelty Detection in Data Streams	133
9.1	Introduction	133
9.2	Learning and Novelty	134
9.2.1	Desiderata for Novelty Detection	135
9.3	Novelty Detection as a One-Class Classification Problem	135
9.3.1	Autoassociator Networks	136
9.3.2	The Positive Naive-Bayes	137
9.3.3	Decision Trees for One-Class Classification	138
9.3.4	The One-Class SVM	138
9.3.5	Evaluation of One-Class Classification Algorithms	139
9.4	Learning New Concepts	141
9.4.1	Approaches Based on Extreme Values	141
9.4.2	Approaches Based on the Decision Structure	142
9.4.3	Approaches Based on Frequency	143
9.4.4	Approaches Based on Distances	144
9.5	The <i>Online Novelty and Drift Detection</i> Algorithm	144
9.5.1	Initial Learning Phase	145

9.5.2	Continuous Unsupervised Learning Phase	146
9.5.2.1	Identifying Novel Concepts	147
9.5.2.2	Attempting to Determine the Nature of New Concepts	149
9.5.2.3	Merging Similar Concepts	149
9.5.2.4	Automatically Adapting the Number of Clus- ters	150
9.5.3	Computational Cost	150
9.6	Notes	151
10	Ensembles of Classifiers	153
10.1	Introduction	153
10.2	Linear Combination of Ensembles	155
10.3	Sampling from a Training Set	156
10.3.1	Online Bagging	157
10.3.2	Online Boosting	158
10.4	Ensembles of Trees	160
10.4.1	Option Trees	160
10.4.2	Forest of Trees	161
10.4.2.1	Generating Forest of Trees	162
10.4.2.2	Classifying Test Examples	162
10.5	Adapting to Drift Using Ensembles of Classifiers	162
10.6	Mining Skewed Data Streams with Ensembles	165
10.7	Notes	166
11	Time Series Data Streams	167
11.1	Introduction to Time Series Analysis	167
11.1.1	Trend	167
11.1.2	Seasonality	169
11.1.3	Stationarity	169
11.2	Time-Series Prediction	169
11.2.1	The Kalman Filter	170
11.2.2	Least Mean Squares	173
11.2.3	Neural Nets and Data Streams	173
11.2.3.1	Stochastic Sequential Learning of Neural Net- works	174
11.2.3.2	Illustrative Example: Load Forecast in Data Streams	175
11.3	Similarity between Time-Series	177
11.3.1	Euclidean Distance	177
11.3.2	Dynamic Time-Warping	178
11.4	Symbolic Approximation – SAX	180
11.4.1	The SAX Transform	180
11.4.1.1	Piecewise Aggregate Approximation (PAA)	181
11.4.1.2	Symbolic Discretization	181

11.4.1.3	Distance Measure	182
11.4.1.4	Discussion	182
11.4.2	Finding <i>Motifs</i> Using SAX	183
11.4.3	Finding <i>Discords</i> Using SAX	183
11.5	Notes	184
12	Ubiquitous Data Mining	185
12.1	Introduction to Ubiquitous Data Mining	185
12.2	Distributed Data Stream Monitoring	186
12.2.1	Distributed Computing of Linear Functions	187
12.2.1.1	A General Algorithm for Computing Linear Functions	188
12.2.2	Computing Sparse Correlation Matrices Efficiently	189
12.2.2.1	Monitoring Sparse Correlation Matrices	191
12.2.2.2	Detecting Significant Correlations	192
12.2.2.3	Dealing with Data Streams	192
12.3	Distributed Clustering	193
12.3.1	Conquering the Divide	193
12.3.1.1	Furthest Point Clustering	193
12.3.1.2	The Parallel Guessing Clustering	193
12.3.2	<i>DGClust</i> – Distributed Grid Clustering	194
12.3.2.1	Local Adaptive Grid	194
12.3.2.2	Frequent State Monitoring	195
12.3.2.3	Centralized Online Clustering	196
12.4	Algorithm Granularity	197
12.4.1	Algorithm Granularity Overview	199
12.4.2	Formalization of Algorithm Granularity	200
12.4.2.1	Algorithm Granularity Procedure	200
12.4.2.2	Algorithm Output Granularity	201
12.5	Notes	203
13	Final Comments	205
13.1	The Next Generation of Knowledge Discovery	205
13.1.1	Mining Spatial Data	206
13.1.2	The Time Situation of Data	206
13.1.3	Structured Data	206
13.2	Where We Want to Go	206
Appendix A	Resources	209
A.1	Software	209
A.2	Datasets	209
Bibliography		211
Index		235