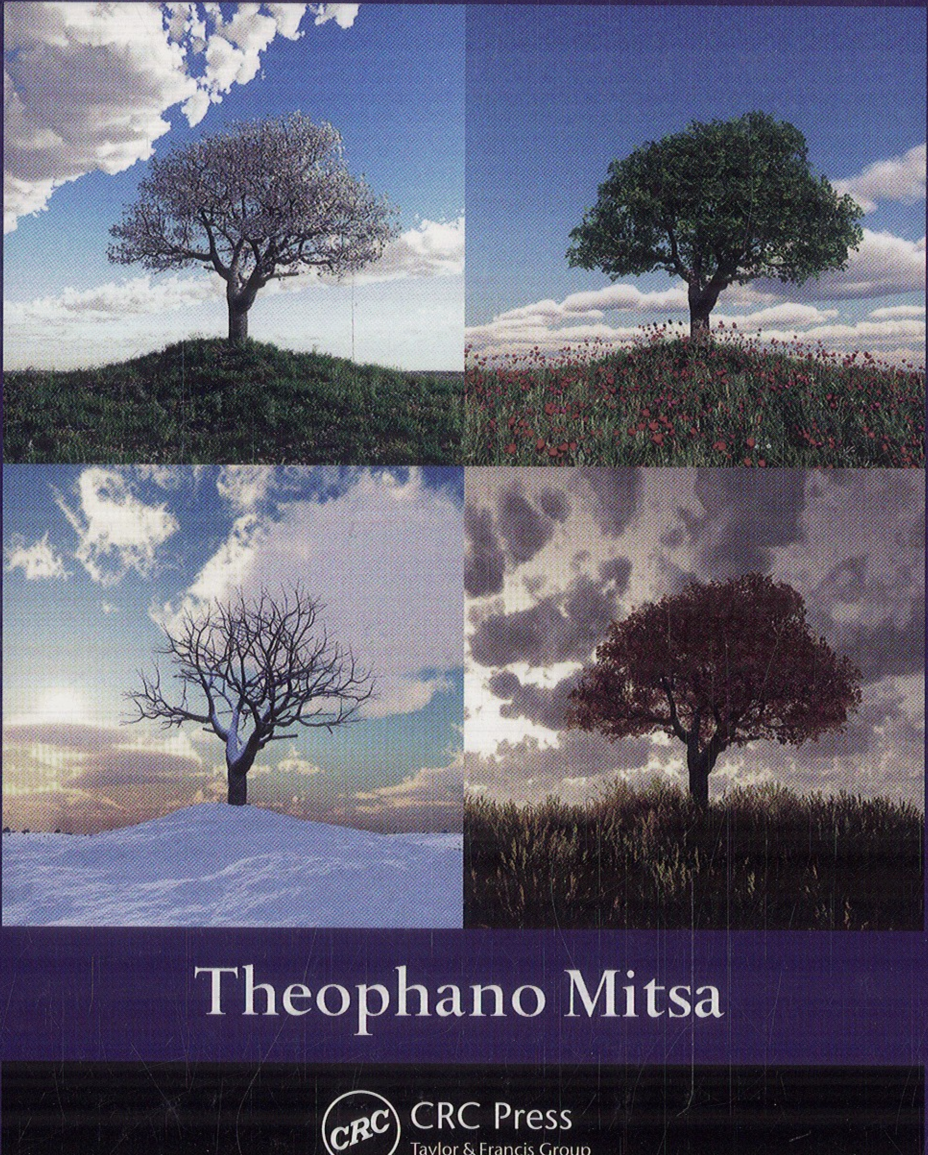
Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series

Temporal Data Mining



Theophano Mitsa

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Table of Contents

Preface, xix

Chapter 1 ▪ Temporal Databases and Mediators 1

1.1 TIME IN DATABASES 1

1.1.1 Database Concepts 2

1.1.2 Temporal Databases 3

1.1.3 Time Representation in SQL 4

1.1.4 Time in Data Warehouses 5

1.1.5 Temporal Constraints and Temporal Relations 5

1.1.6 Requirements for a Temporal Knowledge-Based Management System 6

1.1.7 Using XML for Temporal Data 7

1.1.8 Temporal Entity Relationship Models 8

1.2 DATABASE MEDIATORS 9

1.2.1 Temporal Relation Discovery 10

1.2.2 Semantic Queries on Temporal Data 12

1.3 ADDITIONAL BIBLIOGRAPHY 15

1.3.1 Additional Bibliography on Temporal Primitives 15

1.3.2 Additional Bibliography on Temporal Constraints and Logic 15

1.3.3 Additional Bibliography on Temporal Languages and Frameworks 16

REFERENCES 17

Chapter 2 ▪ Temporal Data Similarity Computation, Representation, and Summarization 21

2.1 TEMPORAL DATA TYPES AND PREPROCESSING 22

2.1.1 Temporal Data Types 22

2.1.2 Temporal Data Preprocessing 22

2.1.2.1 Data Cleaning 22

2.1.2.2 Data Normalization 25

2.2 TIME SERIES SIMILARITY MEASURES 26

2.2.1 Distance-Based Similarity 27

2.2.1.1 Euclidean Distance 27

2.2.1.2 Absolute Difference 28

2.2.1.3 Maximum Distance Metric 28

2.2.2 Dynamic Time Warping 28

2.2.3 The Longest Common Subsequence 31

2.2.4 Other Time Series Similarity Metrics 31

2.3 TIME SERIES REPRESENTATION 33

2.3.1 Nonadaptive Representation Methods 33

2.3.1.1 Discrete Fourier Transform 34

2.3.1.2 Discrete Wavelet Transform 34

2.3.1.3 Piecewise Aggregate Composition 37

2.3.2 Data-Adaptive Representation Methods 38

2.3.2.1 Singular Value Decomposition of Time Sequences 38

2.3.2.2 Shape Definition Language and CAPSUL 39

2.3.2.3 Landmark-Based Representation 40

2.3.2.4 Symbolic Aggregate Approximation (SAX) and iSAX 42

2.3.2.5 Adaptive Piecewise Constant Approximation (APCA) 43

2.3.2.6 Piecewise Linear Representation (PLA) 43

2.3.3 Model-Based Representation Methods 44

2.3.3.1 Markov Models for Representation and Analysis of Time Series 44

2.3.4 Data Dictated Representation Methods 45

2.3.4.1 Clipping 45

2.3.5 Comparison of Representation Schemes and Distance Measures 45

2.3.6 Need for Time Series Data Mining Benchmarks 46

2.4 TIME SERIES SUMMARIZATION METHODS 46

2.4.1 Statistics-Based Summarization 47

2.4.1.1 Mean 47

2.4.1.2 Median 47

2.4.1.3 Mode 47

2.4.1.4 Variance 47

2.4.2 Fractal Dimension–Based Summarization 48

2.4.3 Run-Length–Based Signature 48

2.4.3.1 Short Run-Length Emphasis 49

2.4.3.2 Long Run-Length Emphasis 49

2.4.4 Histogram-Based Signature and Statistical Measures 50

2.4.5 Local Trend-Based Summarization 51

2.5 TEMPORAL EVENT REPRESENTATION 52

2.5.1 Event Representation Using Markov Models 52

2.5.2 A Formalism for Temporal Objects and Repetitions 53

2.6 SIMILARITY COMPUTATION OF SEMANTIC TEMPORAL OBJECTS 54

2.7 TEMPORAL KNOWLEDGE REPRESENTATION IN CASE-BASED REASONING SYSTEMS 55

2.8 ADDITIONAL BIBLIOGRAPHY 56

2.8.1 Similarity Measures 56

2.8.2 Dimensionality Reduction 57

2.8.3 Representation and Summarization Techniques 58
2.8.4 Similarity and Query of Data Streams 59
REFERENCES 59

Chapter 3 ▪ Temporal Data Classification and Clustering 67

3.1 CLASSIFICATION TECHNIQUES 68
3.1.1 Distance-Based Classifiers 68
3.1.1.1 K–Nearest Neighbors 69
3.1.1.2 Exemplar-Based Nearest Neighbor 72
3.1.2 Bayes Classifier 72
3.1.3 Decision Trees 78
3.1.4 Support Vector Machines in Classification 81
3.1.5 Neural Networks in Classification 82
3.1.6 Classification Issues 83
3.1.6.1 Classification Error Types 83
3.1.6.2 Classifier Success Measures 84
3.1.6.3 Generation of the Testing and Training Sets 85
3.1.6.4 Comparison of Classification Approaches 85
3.1.6.5 Feature Processing 85
3.1.6.6 Feature Selection 86
3.2 CLUSTERING 86
3.2.1 Clustering via Partitioning 87
3.2.1.1 K-Means Clustering 87
3.2.1.2 K-Medoids Clustering 88
3.2.2 Hierarchical Clustering 90
3.2.2.1 The COBWEB Algorithm 92
3.2.2.2 The BIRCH Algorithm 92
3.2.2.3 The CURE Algorithm 93
3.2.3 Density-Based Clustering 93

3.2.3.1 The DBSCAN *Algorithm* 94

3.2.4 Fuzzy C-Means Clustering 95

3.2.5 Clustering via the *EM* Algorithm 96

3.3 OUTLIER ANALYSIS AND MEASURES OF CLUSTER VALIDITY 96

3.4 TIME SERIES CLASSIFICATION AND CLUSTERING TECHNIQUES 99

3.4.1 1-NN Time Series Classification 99

3.4.2 Improvement to the *1NN-DTW* Algorithm Using Numerosity Reduction 100

3.4.3 Semi-Supervised Time Series Classification 100

3.4.4 Time Series Classification Using Learned Constraints 101

3.4.5 Entropy-Based Time Series Classification 102

3.4.6 Incremental Iterative Clustering of Time Series 103

3.4.7 Motion Time Series Clustering Using Hidden Markov Models (HMMs) 103

3.4.8 Distance Measures for Effective Clustering of ARIMA Time Series 104

3.4.9 Clustering of Time Series Subsequences 104

3.4.10 Clustering of Time Series Data Streams 105

3.4.11 Model-Based Time Series Clustering 107

3.4.12 Time Series Clustering Using Global Characteristics 107

3.5 ADDITIONAL BIBLIOGRAPHY 108

3.5.1 General Classification and Clustering 108

3.5.2 Time Series/Sequence Classification 108

3.5.3 Time Series Clustering 110

REFERENCES 112

CHAPTER 4 ▪ Prediction 121

4.1 FORECASTING MODEL AND ERROR MEASURES 122

4.2 EVENT PREDICTION 124

4.2.1 Simple Linear Regression 124

4.2.2 Linear Multiple Regression 126

4.2.3 Other Regression Issues 129

4.2.4 Learning to Predict Rare Events in Event Sequences 131

4.3 TIME SERIES FORECASTING 133

4.3.1 Moving Averages 133

4.3.2 Exponential Smoothing 134

4.3.3 Time Series Forecasting via Regression 137

4.3.4 Forecasting Seasonal Data via Regression 137

4.3.5 Random Walk 138

4.3.6 Autocorrelation 140

4.3.7 Autoregression 141

4.3.8 ARMA Models 142

4.4 ADVANCED TIME SERIES FORECASTING TECHNIQUES 143

4.4.1 Neural Networks and Genetic Algorithms in Time Series Forecasting 143

4.4.2 Application of Clustering in Time Series Forecasting 145

4.4.3 Characterization and Prediction of Complex Time Series Events Using Time-Delayed Embedding 146

4.5 ADDITIONAL BIBLIOGRAPHY 147

REFERENCES 149

CHAPTER 5 ▪ Temporal Pattern Discovery 153

5.1 SEQUENCE MINING 154

5.1.1 *Apriori* Algorithm and Its Extension to Sequence Mining 154

5.1.2 The *GSP* Algorithm 157

5.1.2.1 Candidate Generation 158

5.1.3 The *SPADE* Algorithm 159

5.1.4 The *PrefixSpan* and *CloSpan* Algorithms 160
5.1.5 The *SPAM* and *I-SPAM* Algorithms 161
5.1.6 The Frequent Pattern Tree (FP-Tree) Algorithm 162
5.1.7 The *Datte* Algorithm 163
5.1.8 Incremental Mining of Databases for Frequent Sequence Discovery 164
5.2 FREQUENT EPISODE DISCOVERY 165
5.3 TEMPORAL ASSOCIATION RULE DISCOVERY 166
5.3.1 Temporal Association Rule Discovery Using Genetic Programming and Specialized Hardware 167
5.3.2 Meta-Mining of Temporal Data Sets 168
5.3.3 Other Techniques for the Discovery of Temporal Association Rules 168
5.4 PATTERN DISCOVERY IN TIME SERIES 169
5.4.1 Motif Discovery 169
5.4.1.1 General Concepts 169
5.4.1.2 Probabilistic Discovery of Time Series Motifs 170
5.4.1.3 Discovering Motifs in Multivariate Time Series 171
5.4.1.4 Activity Discovery 171
5.4.2 Anomaly Discovery 172
5.4.2.1 General Concepts 172
5.4.2.2 Time Series Discords 172
5.4.2.3 VizTree 173
5.4.2.4 Spacecraft Anomaly Detection Using Support Vector Machines 174
5.4.3 Additional Work in Motif and Anomaly Discovery 175
5.4.4 Full and Partial Periodicity Detection in Time Series 175
5.4.5 Complex Temporal Pattern Identification 178

5.4.6 Retrieval of Relative Temporal Patterns Using Signatures 179

5.4.7 Hidden Markov Models for Temporal Pattern Discovery 179

5.5 FINDING PATTERNS IN STREAMING TIME SERIES 180

5.5.1 *SPIRIT*, *BRAID*, *Statstream*, and Other Stream Pattern Discovery Algorithms 180

5.5.2 Multiple Regression of Streaming Data 181

5.5.3 A Warping Distance for Streaming Time Series 182

5.5.4 Burst Detection in Data Streams 182

5.5.5 The *MUSCLES* and *Selective MUSCLES* Algorithms 183

5.5.6 The *AWSOM* Algorithm 184

5.6 MINING TEMPORAL PATTERNS IN MULTIMEDIA 185

5.7 ADDITIONAL BIBLIOGRAPHY 187

5.7.1 Sequential Pattern Mining 187

5.7.2 Time Series Pattern Discovery 188

REFERENCES 192

CHAPTER 6 ▪ Temporal Data Mining in Medicine and Bioinformatics 201

6.1 TEMPORAL PATTERN DISCOVERY, CLASSIFICATION, AND CLUSTERING 201

6.1.1 Temporal Mining in Clinical Databases 201

6.1.2 Various Physiological Signal Temporal Mining 204

6.1.3 ECG Analysis 208

6.1.4 Analysis and Classification of EEG Time Series 209

6.1.5 Analysis and Clustering of fMRI Data 210

6.1.6 Fuzzy Temporal Data Mining and Reasoning 211

6.1.7 Analysis of Gene Expression Profile Data 212

6.1.7.1 Pattern Discovery in Gene Sequences 213

6.1.7.2 Clustering of Static Gene Expression Data 216

6.1.7.3 Clustering of Gene Expression Time Series 217

6.1.7.4 *Additional Temporal Data Mining–Related Work for Genomic Data* 223

6.1.8 Temporal Patterns Extracted via Case-Based Reasoning 225

6.1.9 Integrated Environments for the Extraction, Processing, and Visualization of Temporal Medical Information 226

6.2 TEMPORAL DATABASES/MEDIATORS 228

6.2.1 Medical Temporal Reasoning 228

6.2.2 Knowledge-Based Temporal Abstraction in Clinical Domains 229

6.2.3 Temporal Database Mediators and Architectures for Abstract Temporal Queries 231

6.2.4 Temporality of Narrative Clinical Information and Clinical Discharge Documents 234

6.2.5 Temporality Incorporation and Temporal Data Mining in Electronic Health Records 235

6.2.6 The BioJournal Monitor 237

6.3 TEMPORALITY IN CLINICAL WORKFLOWS 237

6.3.1 Clinical Workflow Management 237

6.3.2 Querying Clinical Workflows by Temporal Similarity 239

6.3.3 Surgical Workflow Temporal Modeling 240

6.4 ADDITIONAL BIBLIOGRAPHY 240

REFERENCES 243

Chapter 7 ▪ Temporal Data Mining and Forecasting in Business and Industrial Applications 257

7.1 TEMPORAL DATA MINING APPLICATIONS IN ENHANCEMENT OF BUSINESS AND CUSTOMER RELATIONSHIPS 258

7.1.1 Event-Based Marketing and Business Strategy 258

7.1.2 Business Strategy Implementation via Temporal Data Mining 260

7.1.3 Temporality of Business Decision Making and Integration of Temporal Research in Business 261
7.1.4 Intertemporal Economies of Scope 263
7.1.5 Time-Based Competition 264
7.1.6 A Model for Customer Lifetime Value 265
7.2 BUSINESS PROCESS APPLICATIONS 267
7.2.1 Business Process Workflow Management 267
7.2.2 Temporal Data Mining to Measure Operations Performance 267
7.2.3 Temporality in the Supply Chain Management 268
7.2.4 Temporal Data Mining for the Optimization of the Value Chain Management 269
7.2.5 Resource Demand Forecasting Using Sequence Clustering 270
7.2.6 A Temporal Model to Measure the Performance of an IT Project 271
7.2.7 Real-Time Business Analytics 272
7.2.8 Choreographing Web Services for Real-Time Data Mining 272
7.2.9 Temporal Business Rules to Synthesize Composition of Web Services 273
7.3 MISCELLANEOUS INDUSTRIAL APPLICATIONS 273
7.3.1 Temporal Management of RFID Data 273
7.3.2 Time Correlations of Data Streams and Their Effects on Business Impact Analysis 275
7.3.3 Temporal Data Mining in a Large Utility Company 276
7.3.4 The Partition Decoupling Method for Time-Dependent Complex Data 277
7.4 FINANCIAL DATA FORECASTING 277
7.4.1 A Model for Multirelational Data Mining on Demand Forecasting 277
7.4.2 Simultaneous Prediction of Multiple Financial Time Series Using Supervised Learning and Chaos Theory 278

7.4.3 Financial Forecasting through Evolutionary Algorithms and Neural Networks 279

7.4.4 Independent Component Analysis for Financial Time Series 282

7.4.5 Subsequence Matching of Financial Streams 282

7.4.6 Detection of Outliers in Financial Data 283

7.4.7 Stock Portfolio Diversification Using the Fractal Dimension 284

7.5 ADDITIONAL BIBLIOGRAPHY 284

REFERENCES 286

CHAPTER 8 ▪ Web Usage Mining 293

8.1 GENERAL CONCEPTS 293

8.1.1 Preprocessing 294

8.1.2 Pattern Discovery and Analysis in Web Usage 295

8.1.3 Business Applications of Web Usage Mining 296

8.2 WEB USAGE MINING ALGORITHMS 297

8.2.1 Mining Web Usage Patterns 297

8.2.2 Automatic Personalization of a Web Site 298

8.2.3 Measuring and Improving the Success of Web Sites 300

8.2.4 Identification of Online Communities 303

8.2.5 Web Usage Mining in Real Time 304

8.2.6 Mining Evolving User Profiles 304

8.2.7 Identifying Similarities, Periodicities, and Bursts in Online Search Queries 305

8.2.8 Event Detection from Web-Click-Through Data 307

8.3 ADDITIONAL BIBLIOGRAPHY 308

8.3.1 Pattern Discovery 308

8.3.2 Web Usage Mining for Business Applications 309

REFERENCES 310

CHAPTER 9 ▪ Spatiotemporal Data Mining 315

9.1 GENERAL CONCEPTS 315

9.2 FINDING PERIODIC PATTERNS IN SPATIOTEMPORAL DATA 316

9.3 MINING ASSOCIATION RULES IN SPATIOTEMPORAL DATA 317

9.4 APPLICATIONS OF SPATIOTEMPORAL DATA MINING IN GEOGRAPHY 318

9.5 SPATIOTEMPORAL DATA MINING OF TRAFFIC DATA 320

9.6 SPATIOTEMPORAL DATA REDUCTION 321

9.7 SPATIOTEMPORAL DATA QUERIES 322

9.8 INDEXING SPATIOTEMPORAL DATA WAREHOUSES 322

9.9 SEMANTIC REPRESENTATION OF SPATIOTEMPORAL DATA 323

9.10 HISTORICAL SPATIOTEMPORAL AGGREGATION 324

9.11 SPATIOTEMPORAL RULE MINING FOR LOCATION-BASED AWARE SYSTEMS 325

9.12 TRAJECTORY DATA MINING 326

9.13 THE *FLOWMINER* ALGORITHM 327

9.14 THE *TOPOLOGYMINER* ALGORITHM 329

9.15 APPLICATIONS OF TEMPORAL DATA MINING IN THE ENVIRONMENTAL SCIENCES 329

9.16 ADDITIONAL BIBLIOGRAPHY 332

9.16.1 Modeling of Spatiotemporal Data and Query Languages 333

9.16.2 Moving Object Databases 333

REFERENCES 334

APPENDIX A 339

APPENDIX B 345

INDEX 353