MORGAN & CLAYPOOL PUBLISHERS

Outlier Detection for Temporal Data

Manish Gupta
Jing Gao
Charu Aggarwal
Jiawei Han

Synthesis Lectures on Data Mining and Knowledge Discovery

Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, Robert Grossman, *Series Editors*

Outlier Detection for Temporal Data

Synthesis Lectures on Data Mining and Knowledge Discovery

Synthesis Lectures on Data Mining and Knowledge Discovery is edited by Jiawei Han, Lise Getoor, Wei Wang, Johannes Gehrke, and Robert Grossman. The series publishes 50- to 150-page publications on topics pertaining to data mining, web mining, text mining, and knowledge discovery, including tutorials and case studies. The scope will largely follow the purview of premier computer science conferences, such as KDD. Potential topics include, but not limited to, data mining algorithms, innovative data mining applications, data mining systems, mining text, web and semi-structured data, high performance and parallel/distributed data mining, data mining standards, data mining and knowledge discovery framework and process, data mining foundations, mining data streams and sensor data, mining multi-media data, mining social networks and graph data, mining spatial and temporal data, pre-processing and post-processing in data mining, robust and scalable statistical methods, security, privacy, and adversarial data mining, visual data mining, visual analytics, and data visualization.

Outlier Detection for Temporal Data
Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
2014

Provenance Data in Social Media
Geoffrey Barbier, Zhuo Feng, Pritam Gundecha, and Huan Liu
2013

Graph Mining: Laws, Tools, and Case Studies
D. Chakrabarti and C. Faloutsos
2012

Mining Heterogeneous Information Networks: Principles and Methodologies
Yizhou Sun and Jiawei Han
2012

Privacy in Social Networks
Elena Zheleva, Evimaria Terzi, and Lise Getoor
2012

Community Detection and Mining in Social Media
Lei Tang and Huan Liu
2010

Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions
Giovanni Seni and John F. Elder
2010

Modeling and Data Mining in Blogosphere
Nitin Agarwal and Huan Liu
2009

Outlier Detection for Temporal Data

Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han

www.morganclaypool.com

ISBN: 9781627053754 paperback
ISBN: 9781627053761 ebook

DOI 10.2200/S00573ED1V01Y201403DMK008

A Publication in the Morgan & Claypool Publishers series
SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY

Lecture #8
Series Editors: Jiawei Han, *University of Illinois at Urbana-Champaign*
Lise Getoor, *University of Maryland*
Wei Wang, *University of North Carolina, Chapel Hill*
Johannes Gehrke, *Cornell University*
Robert Grossman, *University of Chicago*
Series ISSN
Print 2151-0067 Electronic 2151-0075

Outlier Detection for Temporal Data

Manish Gupta
Microsoft, India and International Institute of Technology–Hyderabad, India

Jing Gao
State University of New York, Buffalo, NY

Charu Aggarwal
IBM T. J. Watson Research Center, NY

Jiawei Han
University of Illinois at Urbana-Champaign, IL

SYNTHESIS LECTURES ON DATA MINING AND KNOWLEDGE DISCOVERY #8

MORGAN & CLAYPOOL PUBLISHERS

ABSTRACT

Outlier (or anomaly) detection is a very broad field which has been studied in the context of a large number of research areas like statistics, data mining, sensor networks, environmental science, distributed systems, spatio-temporal mining, etc. Initial research in outlier detection focused on time series-based outliers (in statistics). Since then, outlier detection has been studied on a large variety of data types including high-dimensional data, uncertain data, stream data, network data, time series data, spatial data, and spatio-temporal data. While there have been many tutorials and surveys for general outlier detection, we focus on outlier detection for temporal data in this book.

A large number of applications generate temporal datasets. For example, in our everyday life, various kinds of records like credit, personnel, financial, judicial, medical, etc., are all temporal. This stresses the need for an organized and detailed study of outliers with respect to such temporal data. In the past decade, there has been a lot of research on various forms of temporal data including consecutive data snapshots, series of data snapshots and data streams. Besides the initial work on time series, researchers have focused on rich forms of data including multiple data streams, spatio-temporal data, network data, community distribution data, etc.

Compared to general outlier detection, techniques for temporal outlier detection are very different. In this book, we will present an organized picture of both recent and past research in temporal outlier detection. We start with the basics and then ramp up the reader to the main ideas in state-of-the-art outlier detection techniques. We motivate the importance of temporal outlier detection and brief the challenges beyond usual outlier detection. Then, we list down a taxonomy of proposed techniques for temporal outlier detection. Such techniques broadly include statistical techniques (like AR models, Markov models, histograms, neural networks), distance- and density-based approaches, grouping-based approaches (clustering, community detection), network-based approaches, and spatio-temporal outlier detection approaches. We summarize by presenting a wide collection of applications where temporal outlier detection techniques have been applied to discover interesting outliers.

KEYWORDS

temporal outlier detection, time series data, data streams, distributed data streams, temporal networks, spatiotemporal outliers

To my dear parents, Satyapal Gupta and Madhubala Gupta,
and my cute loving wife Nidhi

–Manish Gupta

To my husband Lu,
and my parents

–Jing Gao

To my wife Lata,
and my daughter Sayani

–Charu Aggarwal

To my wife Dora,
and my son Lawrence

–Jiawei Han

Contents

Preface xiii

Acknowledgments xv

Figure Credits xvii

1 Introduction and Challenges 1

1.1 Temporal Outlier Examples 2
1.2 Different Facets of Temporal Outlier Analysis 3
1.3 Specific Challenges for Outlier Detection for Temporal Data 4
1.4 Conclusions and Summary 5

2 Outlier Detection for Time Series and Data Sequences 7

2.1 Outliers in Time Series Databases 7
2.1.1 Direct Detection of Outlier Time Series 7
2.1.2 Window-Based Detection of Outlier Time Series 14
2.1.3 Outlier Subsequences in a Test Time Series 16
2.1.4 Outlier Points across Multiple Time Series 16
2.2 Outliers Within a Given Time Series 17
2.2.1 Points as Outliers 17
2.2.2 Subsequences as Outliers 18
2.3 Conclusions and Summary 20

3 Outlier Detection for Data Streams 21

3.1 Evolving Prediction Models 21
3.1.1 Online Sequential Discounting 22
3.1.2 Dynamic Cluster Maintenance 24
3.1.3 Dynamic Bayesian Networks (DBNs) 27
3.2 Distance-Based Outliers for Sliding Windows 29
3.2.1 Distance-Based Global Outliers 30
3.2.2 Distance-Based Local Outliers 31
3.3 Outliers in High-dimensional Data Streams 32

3.4 Detecting Aggregate Windows of Change . . . 33
3.5 Supervised Methods for Streaming Outlier Detection . . . 36
3.6 Conclusions and Summary . . . 36

4 Outlier Detection for Distributed Data Streams . . . 39

4.1 Examples and Challenges . . . 39
4.2 Sharing Data Points . . . 41
4.3 Sharing Local Outliers and Other Data Points . . . 42
4.4 Sharing Model Parameters . . . 43
4.5 Sharing Local Outliers and Data Distributions . . . 45
4.6 Vertically Partitioned Distributed Data . . . 47
4.7 Conclusions and Summary . . . 48

5 Outlier Detection for Spatio-Temporal Data . . . 49

5.1 Spatio-Temporal Outliers (ST-Outliers) . . . 49
5.1.1 Density-Based Outlier Detection . . . 50
5.1.2 Outlier Detection using Spatial Scaling . . . 50
5.1.3 Outlier Detection using Voronoi Diagrams . . . 52
5.2 Spatio-Temporal Outlier Solids . . . 52
5.2.1 Using Kulldorff Scan Statistic . . . 52
5.2.2 Using Image Processing . . . 54
5.3 Trajectory Outliers . . . 55
5.3.1 Distance Between Trajectories . . . 55
5.3.2 Direction and Density of Trajectories . . . 56
5.3.3 Historical Similarity . . . 57
5.3.4 Trajectory Motifs . . . 58
5.4 Conclusions and Summary . . . 59

6 Outlier Detection for Temporal Network Data . . . 61

6.1 Outlier Graphs from Graph Time Series . . . 61
6.1.1 Weight Independent Metrics . . . 62
6.1.2 Metrics using Edge Weights . . . 63
6.1.3 Metrics using Vertex Weights . . . 65
6.1.4 Scan Statistics . . . 67
6.2 Multi-Level Outlier Detection from Graph Snapshots . . . 67
6.2.1 Elbows, Broken Correlations, Prolonged Spikes, and Lightweight Stars . . . 68

6.2.2 Outlier Node Pairs 70
6.3 Community-Based Outlier Detection Algorithms 71
6.3.1 Community Outliers using Community Change Patterns 72
6.3.2 Change Detection using Minimum Description Length 73
6.3.3 Community Outliers using Evolutionary Clustering 73
6.4 Online Graph Outlier Detection Algorithms 74
6.4.1 Spectral Methods 75
6.4.2 Structural Outlier Detection 75
6.5 Conclusions and Summary 75

7 Applications of Outlier Detection for Temporal Data 77
7.1 Temporal Outliers in Environmental Sensor Data 77
7.2 Temporal Outliers in Industrial Sensor Data 80
7.3 Temporal Outliers in Surveillance and Trajectory Data 81
7.4 Temporal Outliers in Computer Networks Data 82
7.5 Temporal Outliers in Biological Data 83
7.6 Temporal Outliers in Astronomy Data 84
7.7 Temporal Outliers in Web Data 84
7.8 Temporal Outliers in Information Network Data 84
7.9 Temporal Outliers in Economics Time Series Data 85
7.10 Conclusions and Summary 86

8 Conclusions and Research Directions 87

Bibliography 91

Authors' Biographies 109

Preface

Temporal data is omnipresent and growing rapidly. Given such huge amounts of temporal data, an important task is to find surprising instances efficiently. Recently, many effective and efficient temporal anomaly detection techniques have been proposed in a variety of research disciplines including data mining, sensor networks, environmental science, distributed systems, spatio-temporal mining, etc. Although there have been multiple surveys and books on general outlier detection, there is no single survey or book dedicated to a thorough study of the diverse techniques and extensive studies in temporal outlier detection. We believe that an organized and extensive coverage of work from multiple disciplines in this book will greatly benefit researchers in these disciplines and motivate cross fertilization of ideas. We begin by motivating the importance of temporal outlier detection and briefing the challenges beyond usual outlier detection. Then, we list down a taxonomy of proposed techniques for temporal outlier detection. For each temporal data type, we will list down several interesting outlier definitions and present various approaches for efficient and effective detection of such outliers. We summarize by presenting a collection of applications where temporal outlier detection techniques have been applied to discover interesting outliers.

SCOPE OF THIS BOOK

This book covers outlier detection techniques for temporal data popular in the data mining community. Many techniques have also been developed in the statistics community and we will not cover them. Specifically, we will discuss techniques for time series data, data streams, distributed data streams, network data, and spatio-temporal data. We will not cover novelty detection techniques.

DEVELOPMENT OF THE BOOK

Many tutorials dedicated to outlier detection were conducted by researchers in data mining, sensor networks, communication networks, and distributed systems communities. Outlier detection is so popular and useful for industry that many tools have been built for efficient outlier detection. For example, the package "outliers" in R, RapidMiner, Oracle, etc. Besides these, many workshops also focused on the general area of outlier detection. However, all of these events focused on the general area of outlier detection; none of these have focused specifically on temporal outlier detection.

This book is based on three tutorials offered by the authors at CIKM 2013, SDM 2013, and ASONAM 2013. A short version of this book also appeared as a survey paper recently published

at TKDE 2014. The networks part of this book draws significant amount of material from the Ph.D. thesis of the first author.

AUDIENCE

This book is mainly targeted at researchers and practitioners in knowledge management, data mining, distributed systems, and sensor networks. While the audience with a good background on data mining would benefit most from this book, we believe the material would give a general audience and newcomers a complete picture of the current work, introduce important research topics in this field, and inspire them to learn more.

Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
March 2014

Acknowledgments

The work was supported in part by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-11-2-0086 (Cyber-Security) and W911NF-09-2-0053 (NS-CTA), the U.S. Army Research Office under Cooperative Agreement No. W911NF-13-1-0193, and U.S. National Science Foundation grants CNS-0931975, IIS-1017362, and IIS-1320617. The views and conclusions contained in our research publications are those of the authors and should not be interpreted as representing any funding agencies. The support is gratefully acknowledged.

We thank Hui Xiong, Leman Akoglu, and Hanghang Tong for their detailed reviews. We thank Diane Cerra, C.L. Tondo, Sara Kreisman, and other members of the Morgan & Claypool team for their patience and their superb production job. Finally, we wish to thank our families, for their constant support and encouragement.

Manish Gupta, Jing Gao, Charu Aggarwal, and Jiawei Han
March 2014

Figure Credits

Figure 3.4 based on Angiulli, F. and Fassetti, F. (2007). Detecting Distance-based Outliers in Streams of Data. *Proceeding CIKM '07 Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 811-820. Copyright © 2007, Association for Computing Machinery, Inc. DOI: 10.1145/1321440.1321552

Figure 4.2 based on Subramaniam, et al: (2006). Online Outlier Detection in Sensor Data using Non-parametric Models. *Proceeding VLDB '06 Proceedings of the 32nd international conference on Very large data bases*, pages 187-198. Copyright © 2006, Very Large Data Base Endowment Inc.

Figures 7.1, 7.2 based on Hill, D. J. and Minsker, B. S. (2010). Anomaly Detection in Streaming Environmental Sensor Data: A Data-driven Modeling Approach. Environmental Modelling and Software, 25(9): 1014–1022. Copyright © 2010 Published by Elsevier Ltd. DOI: 10.1016/j.envsoft.2009.08.010

Figures 7.3, 7.4 based on Birant, D. and Kut, A. (2006). Spatio-Temporal Outlier Detection in Large Databases. *Journal of Computing and Information Technology (CIT)*, 14(4), pages 291-297. CIT. Journal of Computing and Information Technology is an open access journal.

Figure 7.5 from Cheng, T. and Li, Z. (2006). A Multiscale Approach for Spatio-Temporal Outlier Detection. *Transactions in GIS*, Volume 10, Issue 2, pages 253–263, March 2006. Copyright © 2006 John Wiley & Sons, Inc. DOI: 10.1111/j.1467-9671.2006.00256.x

Figure 7.6 from Lasaponara, R. (2005). On the use of principal component analysis (PCA) for evaluating interannual vegetation anomalies from SPOT/VEGETATION NDVI temporal series. *Ecological Modelling*, Volume 194, Issue 4, 15 April 2006, pages 429–434. Copyright © 2005 Elsevier B.V. Reprinted by permission. DOI: 10.1016/j.ecolmodel.2005.10.035

Figure 7.7 based on Lu, C.-T. and Liang, L. (2004). Wavelet fuzzy classification for detecting and tracking region outliers in meteorological data. *Proceeding GIS '04 Proceedings of the 12th annual ACM international workshop on Geographic information systems*, pages 258-265. Copyright © 2004, Association for Computing Machinery, Inc. DOI: 10.1145/1032222.1032260

Figures 7.8, 7.9 based on Dasgupta, D. and Forrest, S. (1996). Novelty Detection in Time Series Data using Ideas from Immunology. *Proceedings of the 5th International Conference on Intelligent Systems*.

Figures 7.10, 7.11, 7.12, 7.13 from Ge, et al: (2010). Top-Eye: Top-K Evolving Trajectory Outlier Detection. *Proceeding CIKM '10 Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1733-1736. Copyright © 2010, Association for Computing Machinery, Inc. Reprinted by permission. DOI: 10.1145/1871437.1871716

Figure 7.14, 7.15 from Keogh, et al: (2005). HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. *Fifth IEEE International Conference on Data Mining*, pages 226-233. Copyright © 2005 IEEE. Used with permission. DOI: 10.1109/ICDM.2005.79

Figure 7.15 based on Keogh, et al: (2005). HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence. *Fifth IEEE International Conference on Data Mining*, pages 226-233. Copyright © 2005 IEEE. DOI: 10.1109/ICDM.2005.79

Figure 7.16 based on Wei, et al: (2006). SAXually Explicit Images: Finding Unusual Shapes. *Sixth International Conference on Data Mining, 2006. ICDM '06*, pages 711-720. Copyright © 2006 IEEE. DOI: 10.1109/ICDM.2006.138

Figure 7.17 from Wei, et al: (2006). SAXually Explicit Images: Finding Unusual Shapes. *Sixth International Conference on Data Mining, 2006. ICDM '06*, pages 711-720. Copyright © 2006 IEEE. Used with permission. DOI: 10.1109/ICDM.2006.138

Figure 7.19 based on Gupta, et al: (2012). Community Trend Outlier Detection Using Soft Temporal Pattern Mining. *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, pp 692-708. Copyright © 2012, Springer-Verlag Berlin Heidelberg. DOI: 10.1007/978-3-642-33486-3_44

CHAPTER 1

Introduction and Challenges

Outlier detection is a broad field that has been studied in the context of a large number of application domains. [Aggarwal, 2013], [Chandola et al., 2009], [Hodge and Austin, 2004], and [Zhang et al., 2008] provide an extensive overview of outlier detection techniques. Outlier detection is also referred to as anomaly detection, event detection, novelty detection, deviant discovery, change point detection, fault detection, intrusion detection, or misuse detection. The three main types of outliers studied in the literature are point outliers, contextual outliers, and collective outliers. Point outliers are generally studied in the context of multidimensional data types, whereas contextual outliers are studied in dependency-oriented data types such as time-series, discrete sequences, spatial data, and graphs. Clearly, the choice of the data type has a significant impact on the methodology used for outlier analysis. A variety of supervised, semi-supervised and unsupervised techniques have been used for outlier detection. These include classification-based, clustering-based, nearest neighbor-based, density-based, statistical, information theory-based, spectral decomposition-based, visualization-based, depth-based, and signal processing-based techniques. Outlier detection has been studied in a variety of data domains including high-dimensional data [Aggarwal and Yu, 2001], uncertain data [Aggarwal and Yu, 2008], streaming data [Aggarwal and Subbian, 2012], [Aggarwal, 2005a], [Aggarwal et al., 2011], network data [Aggarwal et al., 2011], [Gao et al., 2010], [Ghoting et al., 2004], [Gupta et al., 2012a], [Gupta et al., 2012b], and time series data [Burman and Otto, 1988], [Fox, 1972]. Outlier detection has been used extensively for intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting eco-system disturbances. Outlier detection is very popular in industrial applications, and therefore many software tools have been built for efficient outlier detection, such as R (packages "outliers"[1] and "outlierD" [Cho et al., 2008]), SAS,[2] RapidMiner,[3] and Oracle datamine.[4]

Different kinds of data, such as credit, personnel, financial, judicial, medical, and web usage data are temporal. Social network data streams, astronomy data, sensor data, computer network traffic, and commercial transactions are all examples of massive amounts of temporal data. As a result, over time, besides time series, a large variety of temporal datasets have become quite popular. These include temporal networks, temporal databases, data streams, distributed data streams, and spatio-temporal data. The quest to mine information from these new forms of temporal data

[1] http://cran.r-project.org/web/packages/outliers/outliers.pdf
[2] http://www.nesug.org/Proceedings/nesug10/ad/ad07.pdf
[3] http://www.youtube.com/watch?v=C1KNb1Kw-As
[4] http://docs.oracle.com/cd/B28359_01/datamine.111/b28129/anomalies.htm

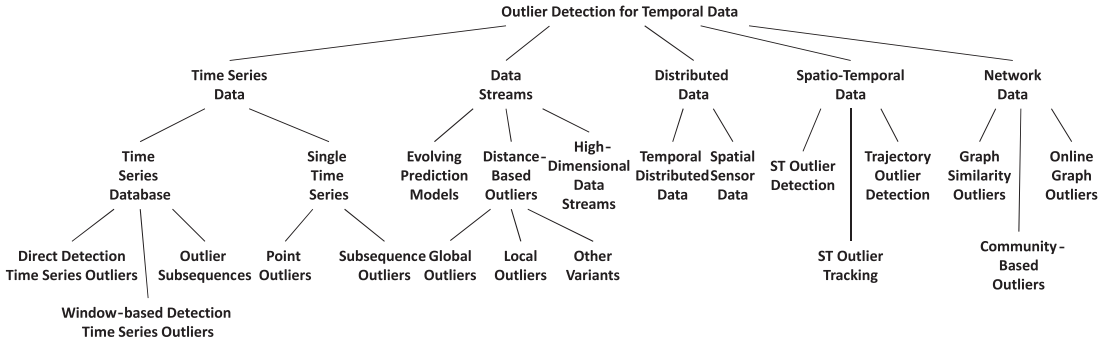

Figure 1.1: Organization of the book.

have led to the rise of the field of temporal data mining [Mitsa, 2010] which includes tasks such as temporal data similarity computation, representation, and summarization, temporal data classification and clustering, prediction, temporal pattern discovery, spatio-temporal data mining, and outlier detection for temporal data. This book focuses on the aspect of outlier detection. A short version of this book has also been published recently [Gupta et al., 2014a] as a journal survey paper. Readers are referred to this survey for a brief overview of the topic.

The different data domains in outlier analysis typically require dedicated techniques of different types. Temporal outlier analysis typically examines anomalies in the behavior of the data *across time*. Therefore the outlierness of a data point can only be understood in the context of the temporal *changes* in the data values or patterns.

1.1 TEMPORAL OUTLIER EXAMPLES

Some real-life examples of outliers in temporal data are as follows.

- *Financial Markets:* An abrupt change in the stock market, or an unusual pattern within a specific window such as the *flash crash* of May 6, 2010, is an anomalous event which needs to be detected early in order to avoid and prevent extensive disruption of markets because of possible weaknesses in trading systems.
- *System Diagnosis:* A significant amount of data generated about the *system state* is discrete in nature. This could correspond to UNIX system calls, aircraft system states, mechanical systems, or host-based intrusion detection systems. The last case is particularly common, and is an important research area in its own right. Anomalies provide information about potentially threatening and failure events in such systems.
- *Biological Data:* While biological data is not temporal in nature, the placement of individual amino-acids is analogous to positions in temporal sequences. Therefore, temporal methods can be directly used for biological data.

- *User-action Sequences:* A variety of sequences abound in daily life that are created by user actions in different domains. These include web browsing patterns, customer transactions, or RFID sequences. Anomalies provide an idea of user-behavior which is deviant for specific reasons (e.g., an attempt to crack a password will contain a sequence of *login* and *password* actions).

This broad diversity in applications is also reflected in the diverse formulations and data types relevant to outlier detection. A common characteristic of all temporal outlier analysis is that *temporal continuity* plays a key role in all these formulations, and unusual *changes*, *sequences*, or *temporal patterns* in the data are used in order to model outliers. In this sense, time forms the *contextual variable* with respect to which all analysis is performed. Temporal outlier analysis is closely related to change point detection, and event detection, since these problems represent two instantiations of a much broader field. The problem of *forecasting* is closely related to many forms of temporal outlier analysis, since outliers are often defined as deviations from *expected* values (or *forecasts*). Nevertheless, while forecasting is a useful *tool* for many forms of outlier analysis, the broader area seems to be much richer and multi-faceted.

1.2 DIFFERENT FACETS OF TEMPORAL OUTLIER ANALYSIS

Outlier analysis problems in temporal data may be categorized in a wide variety of ways that represent different facets of the analysis. The area is so rich that no single type of abstract categorization can fully capture the complexity of the problems in the area, since these different facets may be present in an arbitrary combination. Some of these facets are as follows.

- *Time-series vs. Multidimensional Data Facet:* In time-series data (e.g., sensor readings) the importance of temporal continuity is paramount, and all analysis is performed with careful use of reasonably small windows of time (the contextual variable). On the other hand, in a multi-dimensional data stream such as a text newswire stream, an application such as *first-story detection*, might not rely heavily on the temporal aspect, and thus the methods are much closer to standard multi-dimensional outlier analysis.

- *The Point vs. Window Facet:* Are we looking for an unusual data *point* in a temporal series (e.g., sudden jump in heart rate in ECG reading), or are we looking for an unusual pattern of changes (contiguous ECG pattern indicative of arrythmia)? The latter scenario is usually far more challenging than the former. Even in the context of a multi-dimensional data stream, a single point deviant (e.g., first story in a newswire stream) may be considered a different kind of outlier than an aggregate change point (e.g., sudden change in the aggregate distribution of stories over successive windows).

- *The Data Type Facet:* Different kinds of data such as continuous series (e.g., sensors), discrete series (e.g., web logs), multi-dimensional streams (e.g., text streams), or network data (e.g., graph and social streams) require different kinds of dedicated methods for analysis.

- *The Supervision Facet:* Are previous examples of anomalies available? This facet is of course common to all forms of outlier analysis, and is not specific to the temporal scenario.

These different facets are largely independent of one another, and a large number of problem formulations are possible with the use of a combination of these different facets. Therefore, this book is largely organized by the facet of data type, and examines different kinds of scenarios along this broad organization.

1.3 SPECIFIC CHALLENGES FOR OUTLIER DETECTION FOR TEMPORAL DATA

Compared to the other data mining tasks like classification and clustering, outlier detection presents its unique challenges as follows.

- Classification makes use of available labeled data to learn a classifier model which can then be used to classify future data points. Outlier detection is an unsupervised technique. Outlier detection techniques need to learn similarities between data points without using any user supplied label. Those data points which are very different from others are then marked as outliers. Outlier detection for temporal data aims at identifying anomalous behavior across time. Thus, due to its unsupervised nature, outlier detection becomes challenging.

- Clustering is closely related to outlier detection. Clustering is also an unsupervised technique like outlier detection. Clustering aims at grouping similar objects. Often times, objects that cannot be assigned to any of the clusters are interesting but ignored by the clustering process. These points could be errors in data, noise, or surprising and therefore interesting data points. Outlier detection aims at identifying such surprising data points and is therefore challenging. Temporal clustering aims at maintaining cluster information across time. Outlier detection for temporal data becomes more challenging because it needs to identify data points with surprising combination of temporal properties.

While temporal outlier detection aims to find rare and interesting instances, as in the case of traditional outlier detection, new challenges arise due to the nature of temporal data. We list them below.

- A wide variety of anomaly models are possible depending upon the specific data type and scenario. For example, even though discrete sequences can be viewed as categorical versions of time series, the methods for finding anomalies are quite different in these cases. In temporal graphs, structural patterns need to be accounted for in anomaly detection. This

leads to diverse formulations that need to be designed for the specific problem. For arbitrary applications, it may often not be possible to use off-the-shelf models, because of the wide variations in problem formulations. This is one of the motivating reasons for this book to provide an overview of the most common combinations of facets explored in temporal outlier analysis.

- Since new data arrives at every time instant, the scale of the data is very large. This often leads to processing and resource-constraint challenges. In the streaming scenario, only a single scan is allowed. Traditional outlier detection is much easier, since it is typically an offline task. Unlike the static case, in the dynamic case, an outlier detection system is expected to flag new time points as anomalous or not in real-time. Real-time detection and decision making is a major challenge.
- Outlier detection for temporal data in distributed scenarios poses significant challenges of minimizing communication overhead and computational load in resource-constrained environments.

1.4 CONCLUSIONS AND SUMMARY

Outlier detection is a broad field that has been studied in the context of a large number of application domains. More recently, there has been a large focus on analysis of temporal data and hence a large number of mechanisms for outlier detection for temporal data have been developed. Temporal outliers exist in abundance in real life. Temporal outlier detection is more challenging than the traditional outlier detection because of the scale and the online processing requirement. In this short book, we aim to provide a comprehensive and structured overview of outlier detection techniques for temporal data. Figure 1.1 shows the organization of the book with respect to the data type facet. For each data type, we discuss specific problem classes in various sections. We begin with outlier detection for the easiest scenario for temporal data—discrete time series data in Chapter 2. However, a lot of data gets sampled over very short time intervals, and keeps flowing in infinitely leading to data streams. We study the techniques for outlier detection in streams in Chapter 3. Often times, data is distributed across multiple locations. We study how to extract global outliers in such distributed scenarios in Chapter 4. For some applications like environmental data analysis, data is available over a continuum of both space and time dimensions. We provide an overview of techniques to handle such data in Chapter 5. Finally, networks can capture very rich semantics for almost every domain. Hence, we discuss outlier detection mechanisms for network data in Chapter 6. We also present a few applications where such temporal outlier detection techniques have been successfully employed in Chapter 7. The conclusions are presented in Chapter 8.