Temporal Outlier Analysis

Sunil Kumar Rajwar
Assistant Professor
University Dept. of Computer Applications,
Vinoba Bhave University,
Hazaribag, Jharkhand

I. Mukherjee, PhD
Assistant Professor
Department of Computer Science & Engineering,
Birla Institute of Technology, Mesra, Ranchi,
Jharkhand, India

Pankaj Kumar Manjhi, PhD
Assistant Professor
University Department of Mathematics,
Vinoba Bhave University,
Jharkhand, India

ABSTRACT
The main focus of this research is temporal data. Temporal data means data depends on time. A large number of applications generate a set of temporal data. For example, in our daily life there are different types of records such as credit, personal, financial, judicial, medical, etc. All depends on time. This highlights the need for a detailed organized study of temporal outlier analysis. Over the past decade, a great deal of research has been done on various types of temporal data, including consecutive data snapshots, a series of data snapshots, and data streams. In addition to initial work on time series, the researchers focused on rich data types that include multiple data streams, spatio-temporal data, network data, community distribution data, etc. Compared to general outlier analysis, the techniques for temporal outlier analysis are very different, such as AR models, Markov models, evolutionary clustering, etc.

General Terms
Temporal data , Outlier Detection

Keywords
Temporal Outlier Analysis,Clustering,Time data

1. INTRODUCTION
Data mining, in general, deals with obtaining hidden and interesting information from different types of data. With the development of information technology, the number of databases, as well as their characteristics and complexity, is growing rapidly. What we need is an automated analysis of a lot of information.

The detection of anomalies(or outliers) is a very broad field that has been studied in the context of a large number of research areas such as statistics, data mining, sensor networks, environmental sciences, distribution systems, space-time mining , etc. It has been used for centuries to detect anomalous observations and, where appropriate, to extract anomalous data. changes in system behavior, fraudulent behavior, human error, instrument error, or simply due to natural tendencies in populations. It can identify errors and eliminate their corrupting effect on the data set and thus purify the data for processing. The original outlier detection methods were arbitrary, but now mainstream and systematic techniques are used. Field discovery does this by analyzing and comparing the time series of usage statistics. For application processing, such as loan application processing or social security benefit payments, a scheme detection system can detect any application anomalies before it is approved or paid.

Outline detection can monitor the circumstances of an applicant over time to ensure that the payment is not a fraud. Stock or commodity traders can use anomaly detection methods to monitor individual stocks or markets for new trends that may indicate buying or selling opportunities. A news delivery system can detect changing news and ensure that the provider is the first to receive broken news. Either way, anomaly detection is critical to the consistency and integrity of the database.

2. LITERATURE SURVEY
OUTLIER detection is a broad field, which has been studied in the context of a large number of application domains. Aggarwal [1], Chandola et al. [2], Hodge et al. [3] and Zhang et al. [4] provide a comprehensive description of perimeter detection techniques.

Anomaly detection was studied in a variety of data domains including high-dimensional data [5], indeterminate data [6], data transmission [7], [8], [9], network

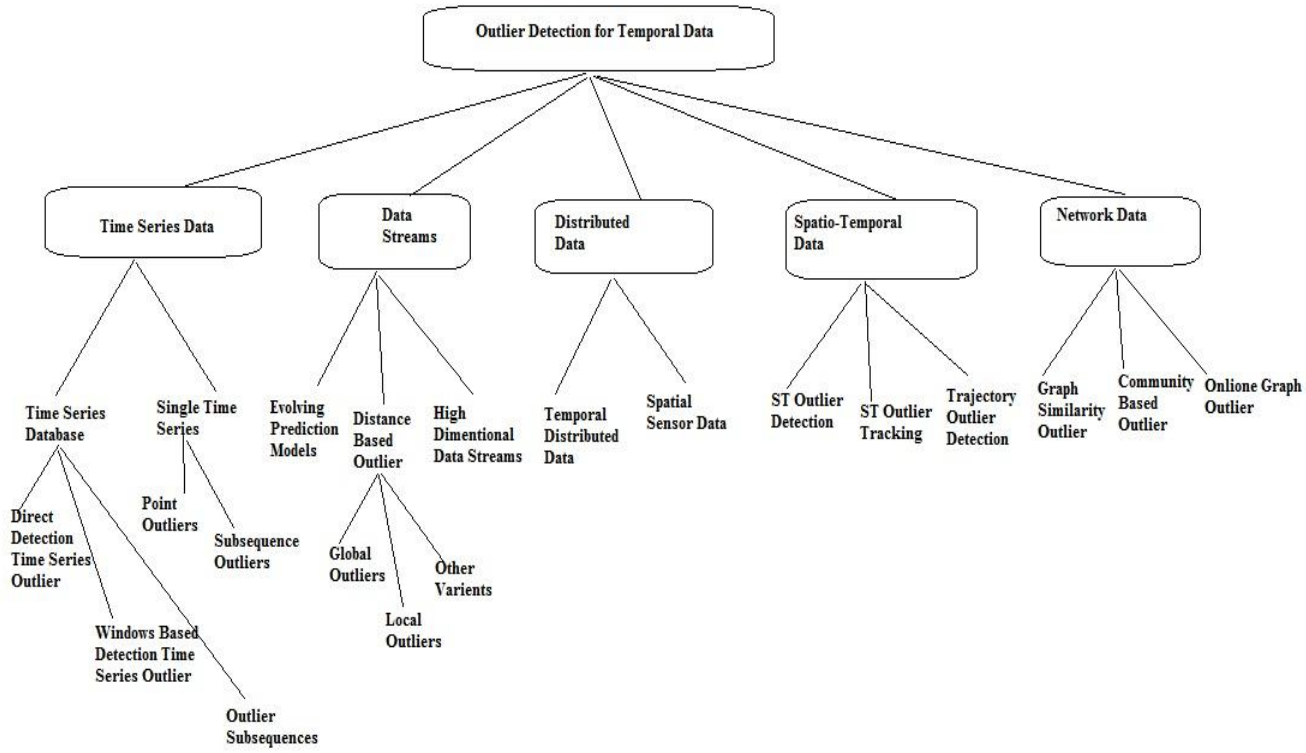
data [9], [10], [11], [12], [13] and time series data [14], [15]. Anomaly detection is very popular in industrial applications and therefore many software tools exist designed for efficient anomaly discovery, such as R ("outliers" and "outlierD" packages [16]), SAS, Rapid-Miner, and Oracle Datamine. Different data domains in schema analysis generally require different types of dedicated techniques. Temporal Outlier analysis examines anomalies in the behavior of data over time. Application domains of Temporal Outlier Analysis are as follows:

Financial markets: A sudden change in the stock market, or an unusual pattern within a particular window, such as the catastrophe of May 6, 2010, is an anomalous event that must be detected early to avoid and prevent widespread market disruption.

System Diagnostics - A significant amount of data generated on the health of the system is isolated by nature. This can respond to calls from UNIX systems, aircraft system states, mechanical systems, or host-based intrusion detection systems. The latter case is particularly common and is an important area of investigation in its own right. Anomalies provide information on potentially threatening events and failures in those systems.

Biological data: Although biological data is not time data, the location of individual amino acids is similar to the sites in time sequences. Therefore, time methods can be used directly for biological data.

User activity sequences: There are different sequences in everyday life, which create user actions in different domains. These include web browsing patterns, customer transactions, or RFID sequences. The anomalies give insight into the behavior of the user deviating for specific reasons (for example, a sequence of login and password actions in an attempt to crack a password).



This wide variety of applications is also reflected in the various formulations and data types associated with Anomaly Detection. A common feature of all Temporal Analysis is that time continuity plays a key role in all of these formulations, and changes, sequences, or time patterns are used in the data to model exemplars. In this sense, time is the contextual variable for which all the analysis is performed.

3. DESIGN AND METHODOLOGY

Temporal Outlier analysis is closely related to point detection and incident detection, as these problems represent two cases of much larger field. The problem of forecasting is closely related to many types of temporal anomaly analysis, as outliers are often defined as deviations from expected (or forecast) values. However, while forecasting is a useful tool for many types of outlier analysis, the wider area appears to be much richer and more diverse.

Among all the various fields based on Temporal outlier analysis, this research work design relies on Anomaly detection of flow data. A data stream is a command sequence of objects X1, .., Xn. The main difference between a traditional database and a data flow management system (DFMS) is that we have unlimited data flows instead of relationships. Applications, such as fraud detection, network flow monitoring, telecommunications, data management, etc., where data access is constant and it is not necessary or practical to store all incoming objects. Traditional data mining methods cannot be applied for efficient data transmission, because these methods are suitable for the environment where the complete dataset is already available and where the algorithm can operate in more than one pass. . A general framework for mining data streams requires a consistent small time per record as well as the minimum memory requirement, using at least one data scan. Because the nature of data flow is infinite, the problem of perimeter mining in data flows is often solved based on certain time intervals, commonly known as windows.

There are two basic approaches to the Anomaly detection problem:

1. **Type 1 -**Model of normality and anomaly. This approach is similar to supervised classification and requires pre-tagged data, tagged as normal or abnormal.

Classifiers are best suited for static data, because the classification must be recreated from the first principles if the distribution of the data changes, unless the system uses an incremental classifier as an evolutionary neural network. Classification algorithms require a good distribution of normal and abnormal data, that is, the data must cover the entire distribution to allow generalization from the classifier. These robust approaches can support Anomaly in the data and generally encourage a limit of normality around the majority of the data, reflecting normal behavior.

2**. Type** 2 - Find out the Anomaly without prior knowledge of the data. This is essentially a learning approach that is similar to unsupervised clustering. The approach processes the data as a static distribution, identifies the most remote points, and marks them as a potential outflow. Type 1 assumes that errors or failures are separate from “normal” data and will therefore appear as outliers. The approach is largely retrospective and is similar to a batch processing system. It requires that all data be available before it is processed and that the data be static. However, once the system has a large enough database with good coverage, it can compare new items with existing data.

There are two commonly used sub-techniques, diagnosis and accommodation (Rousseeuw and Leroy, 1996). An outline diagnostic approach highlights possible peripheral points. Once detected, the system can eliminate these migrants from future processing of the data distribution. Many types of diagnostic approaches remove outlets and add their system

model to the remaining data until no further outflows are detected.

We noted that outline detection methods are derived from three computational domains: statistics (proximity-based, parametric, non-parametric, and semi-parametric), neural networks (supervised and unsupervised), and machine learning.

Statistical model: Statistical approaches were the first algorithms used to detect migrants. Statistical models are usually suitable for at least quantitative sets of true values or data distributions.

Techniques based on proximity: Techniques based on proximity are easy to implement and do not generate preconceived notions about the data distribution model. The computational complexity is directly proportional to the size of the data m and the number of records n. Therefore, methods like nearest k-neighbor (also called sample-based learning) with run time O (n2m) are not feasible for high-dimensional data sets if current time cannot be improved. The CLARANS data mining segmentation algorithm is an optimized derivative of the k-medoids algorithm (Ng and Han, 1994) and can handle the perimeter detection achieved as a by-product of the clustering process. Implements a random but limited heuristic search to get the best cluster by randomly searching for cluster updates.

Machine learning: Much more peripheral detection focused only on the continuous characteristics of the real data value and there was little focus on categorical data. Most statistical and neural approaches require cardinal data or at least prescriptive data to allow calculation of vector distances and no mechanism to process categorical data without any implicit order. The basic algorithms used in machine learning are Decision Tree, BIRCH and DBSCAN, etc.

Hybrid systems: Hybrid is the latest development in Outlier detection systems. Hybrid systems incorporate algorithms from at least two of the previous divisions (statistical, neural, or machine learning methods). Hybridization is used differently to overcome the shortcomings with unique classification algorithms, to exploit the advantages of multiple approaches and overcome their weaknesses, or to use a meta classifier to solve the results of multiple classifiers to handle each scenario.

4. APPLICATIONS OF TEMPORAL OUTLIER ANALYSIS

Environmental sensor data

Used primarily to identify measurement errors in the wind speed data stream to extract the distant perimeter of the rain cover, sea surface temperature, relative humidity, and precipitation.

Computer networks:
Techniques for anomaly detection from time data t have been widely used to detect intrusions [17], [18], [19], [20], [21], [22]. Lakhina et al. [23] use multi-layer source destination flow periods that measure the number of bytes, packets, and flows at the IP level to find anomalies such as point-to-point measurement transfer, denial of service (DOS), denial of service distributed (DDOS attacks), crowd flash (high demand for resources / services), host scan for vulnerable port or network scan for target port, WORM etc.

Economic time series data:
Several economic time sets were studied for anomaly detection. Gupta et al. [12] identify the anomaly based on the unusual change in GDP components (consumption, investment, public spending, and net exports) over time, using time-lapse community perception detection methods

Web data:
In view of the multiple crawls of the web graph, Papadimitriou et al. [24], [25] obtain a trace graph with anomalies. These anomalies refer to failures of web servers that do not allow the crawler to access their content or to hardware / software problems in the search engine infrastructure that can corrupt parts of the data that is reduced.

5. CONCLUSIONS

Modeling time data is a challenging task due to the dynamic nature and complex evolutionary patterns of the data. In the past, a wide range of models have been developed to capture various features for detecting time data outliers. Although the number of formulations on the temporary anomaly detection problem varies, they are usually inspired by the most common applications found in the literature. There are many temporary anomaly detection formulations, which have not been adequately explored. A typical anomaly detection technique takes a standard model or distribution of data and identifies data points deviated from the model as outliers. These techniques are clearly not suitable for online data streams where the complete dataset, due to its unlimited volume, is not available for random access. In addition, the distribution of data in the data streams changes over time, challenging existing outlier detection techniques that adopt a standard data distribution for the entire dataset. In addition, data flows are characterized by uncertainty imposed by greater complexity.

6. REFERENCES

[1] C. C. Aggarwal, Outlier Analysis. Springer, 2013.

[2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly Detection: A Survey,” ACM Computing Surveys, vol. 41, no. 3, pp.15:1–15:58, 2009.

[3] V. J. Hodge and J. Austin, “A Survey of Outlier Detection Methodologies,” Artificial Intelligence Review, vol. 22, no. 2, pp. 85–126, 2004.

[4] Y. Zhang, N. Meratnia, and P. J. M. Havinga, “Outlier Detection Techniques For Wireless Sensor Networks: A Survey,” Centre for Telematics and Information Technology University of Twente, Tech. Rep. TR-CTIT-08-59, Oct 2008.

[5] C. C. Aggarwal and P. S. Yu, “Outlier Detection for High Dimensional Data,” SIGMOD Records, vol. 30, pp. 37–46, May 2001.

[6] C. C. Aggarwal and P. S. Yu, “Outlier Detection with Uncertain Data,” in Proc. Of the 2008 SIAM Intl. Conf. on Data Mining (SDM), 2008, pp. 483–493.

[7] C. Aggarwal and K. Subbian, “Event Detection in Social Streams,” in Proc. of the 12th SIAM Intl. Conf. on Data Mining (SDM), 2012, pp. 624–635.

[8] C. C. Aggarwal, “On Abnormality Detection in Spuriously Populated Data Streams,” in Proc. of the 2005 SIAM Intl. Conf. on Data Mining (SDM), 2005, pp. 80–91.

[9] C. C. Aggarwal, Y. Zhao, and P. S. Yu, “Outlier

Detection in Graph Streams," in Proc. of the 27th Intl. Conf. on Data Engineering (ICDE). IEEE Computer Society, 2011, pp. 399–409.

[10] J. Gao, F. Liang, W. Fan, C. Wang, Y. Sun, and J. Han, "On Community Outliers and their Efficient Detection in Information Networks," in Proc. of the 16th ACM Intl. Conf. on Knowledge Discovery and Data Mining (KDD), 2010, pp. 813–822. GUPTA et al.: OUTLIER DETECTION FOR TEMPORAL DATA: A SURVEY 17

[11] A. Ghoting, M. E. Otey, and S. Parthasarathy, "LOADED: Link-Based Outlier and Anomaly Detection in Evolving Data Sets," in Proc. of the 4th IEEE Intl. Conf. on Data Mining (ICDM), 2004, pp. 387–390.

[12] M. Gupta, J. Gao, Y. Sun, and J. Han, "Community Trend Outlier Detection using Soft Temporal Pattern Mining," in Proc. of the 2012 European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), 2012, pp. 692–708.

[13] M. Gupta, J. Gao, Y. Sun, and J. Han, "Integrating Community Matching and Outlier Detection for Mining Evolutionary Community Outliers," in Proc. of the 18th ACM Intl. Conf. on Knowledge Discovery and Data Mining (KDD), 2012, pp. 859–867.

[14] J. P. Burman and M. C. Otto, "Census Bureau Research Project: Outliers in Time Series," 1988.

[15] A. J. Fox, "Outliers in Time Series," Journal of the Royal Statistical Society. Series B (Methodological), vol. 34, no. 3, pp. 350–363, 1972.

[16] H. Cho, Y. jin Kim, H. J. Jung, S.-W. Lee, and J. W. Lee, "Outlierd: An r package for outlier detection using quantile regression on mass spectrometry data," Bioinformatics, vol. 24, no. 6, pp. 882–884, 2008.

[17] K. Sequeira and M. Zaki, "ADMIT: Anomaly-based Data Mining for Intrusions," in Proc. of the 8th ACM Intl. Conf. on Knowledge Discovery and Data Mining (KDD), 2002, pp. 386–395.

[18] S. A. Hofmeyr, S. Forrest, and A. Somayaji, "Intrusion Detection using Sequences of System Calls," Journal of Computer Security, vol. 6, no. 3, pp. 151–180, Aug 1998.

[19] T. Lane and C. E. Brodley, "An Application of Machine Learning to Anomaly Detection," in Proc. of the 20th National Information Systems Security Conf. (NISSC), 1997, pp. 366–380.

[20] T. Lane and C. E. Brodley, "Temporal Sequence Learning and Data Reduction for Anomaly Detection," in Proc. of the 5th ACM Conf. on Computer and Communications Security (CCS), 1998, pp. 150–158.

[21] F. Angiulli and F. Fassetti, "Detecting Distance-based Outliers in Streams of Data," in Proc. of the 16th ACM Conf. on Information and Knowledge Management (CIKM), 2007, pp. 811–820.

[22] C. Warrender, S. Forrest, and B. Pearlmutter, "Detecting Intrusions using System Calls: Alternative Data Models," in Proc. of the 1999 IEEE Symposium on Security and Privacy, 1999, pp. 133–145.

[23] A. Lakhina, M. Crovella, and C. Diot, "Characterization of Network-wide Anomalies in Traffic Flows," in Proc. of the 4th ACM SIGCOMM Conf. on Internet Measurement (IMC), 2004, pp. 201–206.

[24] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, "Web Graph Similarity for Anomaly Detection," Journal of Internet Services and Applications, vol. 1, no. 1, pp.19–30, 2010.

[25] P. Papadimitriou, A. Dasdan, and H. Garcia-Molina, "Web Graph Similarity for Anomaly Detection," in Proc. of the 17th Intl. Conf. on World Wide Web (WWW), 2008, pp. 1167–1168.