International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey on Outlier Pattern Detection Techniques for Time-Series Data

Archana N.[1], S. S. Pawar[2]

[1]Department of ME Computer Engineering, D.Y.Patil College of Engineering, Akurdi, Savitribai Phule Pune University-411044, Pune, India

[2]Department of ME Computer Engineering, D.Y.Patil College of Engineering, Akurdi, Savitribai Phule Pune University-411044, Pune, India

Abstract: ***Outlier patterns are unusual or surprising patterns that occur rarely, and, thus, have lesser support (frequency of appearance) in the data. Outlier patterns reveal many hidden facts that indicate inconsistency in the data such as fraudulent transactions, network intrusion, change in customer behavior, epidemic and disease severity, intense weather conditions, recession in the economy, etc. Outlier detection has been studied in a variety of data domains including high-dimensional uncertain data, streaming data, network data and time series data. The scope of this survey is limited to time series data. Detecting these outlier patterns rather than other frequent patterns is more important because outlier patterns indicate interesting discrepancies and is crucial for analysis and further decision-making. Outlier values in the data are different from surprising, unusual, or outlier patterns in the data. Outlier Pattern detection in time-ordered sequences discovers in the time series all patterns that exhibit temporal regularities. Considering temporal aspect, interesting outlier patterns can be discovered which otherwise would not have been discovered. The periodicity detection of outlier patterns is to be performed after the detection of these outlier patterns for better analysis of data. Periodic outlier patterns can be found in heart beat pulses, outlier light curves in catalogs of periodic stars, weather data, transactions history, stock price movement, protein and DNA sequences etc. In this paper, the different outlier detection techniques for time series and the existing algorithms in the area are surveyed.***

Keywords: Periodic patterns, time series, pattern mining, outlier pattern, periodicity detection

1. Introduction

A time series is a sequence of observations of well-defined data items obtained through repeated measurements over time**.** Level of employment measured every month can be considered as an example of time series. If the data is being collected irregularly or all at once it is not a time series. Classic time series examples are meteorological data like temperature or rainfall; economic data like stock prices and also medical data. A times series helps in analyzing the effect of cyclical, seasonal and irregular events on the data item being measured. Outlier detection has several applications in time-series data sets such as attack detection in recommender systems, detection of anomalous flight sequences using sensor data from aircrafts, shape anomalies and so on[3].

"Pattern mining" refers to a data mining method that involves finding existing patterns in data. Frequent pattern mining, sequential pattern mining, item-set mining etc. also fall under pattern mining. Detecting of outlier patterns might be more important in many sequences than that of regular, more frequent patterns for analysis. Change in customer behavior, unusual ECG heart beat, surprising patterns in protein sequences etc., represent outlier patterns. The periodicity of these outlier patterns reveals interesting facts.

Outlier values in the data are different from surprising, unusual, or outlier patterns in the data [2]. There are many techniques to find local and global outliers in the data, to find frequent patterns in the data but detecting outlier patterns is different from detection of outlier values or other frequent patterns. For example, in a certain sequence, events *a* and *b* might not be outliers but the pattern *aba* (a certain combination of the events) might be an outlier pattern. Periodicity detection of outlier patterns is an area of research which is yet to be explored in detail. This survey paper also gives a view on periodic outlier detection as in [2].

2. Classification of Outlier Detection Techniques for Time Series Data

The problem of outlier detection for time series data can be viewed in different ways. One way of classification is that outliers can be detected in univariate and multivariate time series which is studied by Deepthi et al. [3]. Anomaly detection for these cases is distinct from traditional anomaly detection. The problem of outlier detection within a given time series data is classified in the following way:

2.1 Detecting Contextual Anomalies in Time Series

Anomalies are the individual instances of the time series which are anomalous in a specific context, but not otherwise. Statistics Community has made a lot of study related to this problem type. Seasonal Temperature value is an example for this type of outlier.

2.2 Detecting Anomalous Subsequence within a Given Time Series

The solution for this kind of problem tries to find an anomalous subsequence with respect to a given long sequence (time series). This problem can be categorized under unsupervised learning environment due to the lack of labeled data for training, but most part of the long sequence (time series) is assumed to be normal. If the anomalous

subsequence is of unit length, this problem is equivalent to finding contextual anomalies in the time series, which is the problem type 1. The anomalous subsequences can also be called as discords. Contiguous ECG pattern indicative of arrhythmia is an example of anomalous subsequence. This anomalous subsequence cannot be detected using point outlier detection techniques because individually values are normal but as a sequence it is abnormal.

2.3 Detecting anomalous time series with respect to a time series database

This problem tries to determine if a test time series is anomalous with respect to a database of training time series. The database under consideration can be of two types. First type consists of only normal time series, which is a semi-supervised setting. The second type consists of unlabeled time series (unsupervised anomaly detection) of both normal and anomalous data. In the second case majority data are considered normal.

2.4 Detecting Periodic Outlier Pattern

This type has been added to the classification given by Deepthi et al. [3]. A less frequent pattern with larger coverage area (having repetitions in larger subsection of the sequence) is called a periodic outlier pattern [2]. Rasheed et al. [2] has provided suitable examples for this problem type.
The pattern $X = xy$ with period $p = 7$ is a better candidate for outlier pattern in the sequence
S = zxyzxyz **xy**zxyzx **xy**zxyzx **xy**zxyzxyzxy
0123456 7890123 4567890 12345678901
than the pattern $X = xy$ with period $p = 2$ in the sequence
S' = zxyzxyz **xyxyxy** zxyzxyzxyzxyzxyzxyz
0123456 789012 3456789012345678901
In S', one way to explain xyxyxy is as a contiguous random noise, while in sequence *S*, it is very clear that the regular periodic pattern *zxy* is disturbed by the regular noise of *xy* showing up at every seventh position.

Gupta et al. [19] have also organized and provided a detailed study of the work done in the area of outlier detection with respect to temporal datasets including time series data sets. The first type includes techniques to detect outliers over a database of time series, whereas the second type deals with outliers within a single time series.

2.5 Outliers in Time Series Databases

In this type, time series database is given. An outlier score for a time series can be computed directly, or by first computing scores for overlapping fixed size windows and then aggregating them. It involves techniques like Direct Detection of Outlier Time Series, Window based Detection of Outlier Time Series and Outlier Subsequences in a Test Time Series. In Direct detection it is assumed that most of the time series in the database are normal while only a few are anomalous. The model is first learnt based on all the time series sequences in the database similar to traditional outlier detection. An outlier score is then computed for each sequence with respect to the model. The model could be supervised or unsupervised depending on the availability of training data. In Window based outlier detection the test sequence is broken into multiple overlapping subsequences (windows). The anomaly score would be computed for each window, and then the anomaly score (AS) for the entire test sequence is computed in terms of that of the individual windows. Window-based techniques perform better localization of anomalies, compared to the techniques that output the entire time series as outliers directly. These techniques need the window length as a parameter. The database maintained can be Normal Pattern Database Approach Negative and Mixed Pattern Database Approaches. The approaches for Outlier Subsequences in a Test Time Series are covered in section III of this paper.

2.6 Outliers within a given Time Series

This type includes finding points (particular elements) and subsequences as outliers in a given time series. Clustering Approaches, Classification Approaches, Prediction models, Profile based models are used to detect points as outliers. Examples of Prediction Models that can be used for point outlier detection are Single-layer linear network predictor (or AR model), Multilayer Perceptron (MLP) predictor, support vector regression etc. The approaches for subsequences as outliers are covered in section III of this paper.

3. Approaches for detecting outlier patterns in time series

The algorithms proposed by researchers for detecting outlier subsequences in time series are given below:

3.1 Base Algorithm

Fu et al. [7] focused on finding the most unusual time series subsequence and proposed a 'base' algorithm based on some properties of the Haar wavelet transformation. The authors have found Time Series Discords Based on Haar Transform. The algorithm was investigated on electrocardiograms (ECGs) and discords were found successfully. Time series discords are subsequences of a longer time series that are maximally different to all the rest of the time series subsequences. Subsequence comparisons are ordered using Haar transform for effective pruning. In this algorithm, all the possible candidate subsequences in outer loop are extracted, the distance to the nearest non-self match for each candidate subsequence is found in inner loop. The candidate pattern which has the largest distance to its nearest non-self match is declared as discord. Various heuristics have been adopted in every step of this algorithm. This algorithm falls under Subsequences as Outliers classification.

3.2 Heuristic Discord Discovery

Keogh et al. [8] proposed a simple algorithm, Heuristic Discord Discovery. This algorithm is 3 to 4 magnitude faster than brute force that was efficiently finding discords. The effectiveness of the discord discovery algorithm is tested by Keogh et al. [8] on 82 different time-series datasets from diverse domains. Three possible heuristic strategies: random, magic and perverse are described which is followed by

Approximation to Magic Heuristics by using SAX (**S**ymbolic **A**ggregate Appro**X**imation). This algorithm falls under Subsequences as Outliers classification.

3.3 Tarzan Algorithm

Keogh et al. [1] define a soft match version where the frequency of pattern p in the database D is defined using the largest number l such that every subsequence of p of length l occurs at least once in D. Surprising patterns are found using Tarzan Algorithm on a dataset that contains the power demand for a Dutch research facility for the entire year of 1997. This algorithm uses suffix tree to efficiently encode the frequency of all observed patterns and allows Markov Model to predict the expected frequency of previously unnoticed patterns. The suffix tree gets constructed first. A Surprise measure can be determined for all the patterns in the new database. The time and space required is linear in the size of the database. Tarzan is not an acronym. The name was given so because the heart of the algorithm depends on comparing two suffix trees. This algorithm falls under Outlier Subsequences in a Test Time Series.

3.4 InfoMiner Algorithm

Yang et al. [4] have presented their so-called InfoMiner algorithm and its variations [5] which discover "surprising periodic patterns." Their "surprise" prioritizes those patterns involving less frequency and more support. Support means matching repetition. A new kind of measurement was introduced here. It was called information, which values the degree of surprise of each occurrence of a pattern. Information treats occurrence as a continuous and monotonically decreasing function of its probability of occurrence. Thus patterns with different probability occurrences are handled easily. This information gain concept can handle most adverse situations caused by the violation of the downward closure property by the information gain measure. Thus it provides an efficient solution to this problem. This algorithm falls under Outlier Subsequences in a Test Time Series.

Besides the above, Chuah et al. [6] proposed an anomaly detection scheme based on time series analysis that will allow the computer to determine whether a stream of real-time sensor data contains any abnormal heartbeats. If anomaly exists, that time series segment will be transmitted via the network to a physician so that he/she can further diagnose the problem and take appropriate actions. The Adaptive Window Based Discord Discovery scheme that the authors designed was motivated by the two schemes Brute Force Discord Discovery (BFDD) and the Heuristic Discord Discovery (HDD) schemes. Lin et al. [9] used subseries join to obtain the similarity relationships among subseries of the time series data. Then the anomaly detection problems can be converted to graph-theoretic problems solvable by existing graph-theoretic algorithms.

4. Periodic Outlier Patterns

The terminologies related to periodic outlier patterns are:

4.1. Periodic Pattern

Periodic pattern is a regular form that repeats itself with a specific period in a give sequence. Periodic patterns can be mined from datasets like biological sequences, continuous and discrete time series data, spatiotemporal data and social networks. Regularly repeating patterns found in a lengthy sequence are called periodic patterns. Periodic analysis is often performed over time-series data which consists of sequences of values or events typically measured at equal time intervals.

Periodic Pattern is observed in different data sets.

4.1.1 Biological Sequences

An important problem in BioInformatics domain is to discover patterns in protein sequences like DNA sequences of human, protein sequences of bacteria helping in vaccine discovery. Core pattern is observed; the remaining partial slots of the pattern allow some mutations. These patterns discovered can undergo mutations leading to further research in biological area.

4.1.2 Time Series Data

The data collected at uniform time interval gaps is called discrete data. Example: Total Sales of a shopping mall is recorded every day. Thus the collection for a week is discrete time series data. Observing this data for a year might give rise to periodic interesting customer behavior. Continuous Time series data collection is where data is collected at every point of time. Heart beat data of a person is examined continuously for clinical depression. The surprising patterns might help in the patient's treatment.

4.1.3 Spatiotemporal Data

Periodicity is a commonly observed in moving objects. People go to workplace every day through more or less same route, birds and animals migration from one place to another also depict yearly periodicity [14].

4.1.4 Social Networks

Social interactions that occur regularly typically correspond to significant yet often infrequent and hard to detect interaction patterns. To identify such regular behavior is known as periodic pattern mining. It includes mining of periodic sub graphs in dynamic social networks [20]. Communication interaction between people and topic discussion in social media also show periodicity. For example, news about different budgets in news sites is delivered periodically. Posting of information in social networks corresponding to different events is done periodically. For example, annual conferences in blogs and websites are done periodically [14].

4.2 Types of Periodic Patterns

Sirisha et al. [14] have classified the periodic patterns into the following categories.

4.2.1 Full and Partial Periodic Patterns

Full periodic pattern is a pattern where every position in the pattern exhibits periodicity. Periodic patterns in which one or

more elements do not exhibit the periodicity are called partial periodic patterns. Sheng et al. [16] detect partial periodicity. Elfeky et al. [13] detect full-cycle periodicity.

4.2.2 Perfect and imperfect periodic patterns

A pattern X is said to satisfy perfect periodicity in sequence S with period p if starting from the first occurrence of X until the end of S, every next occurrence of X exists p positions away from the current occurrence of X.

4.2.3 Synchronous and Asynchronous Periodic Patterns

A pattern which occurs periodically without any misalignment is called as synchronous periodic pattern. Asynchronous periodic patterns are patterns with some disturbance between the repetitions of the pattern.

4.2.4 Patterns with Symbol, Sequence and Segment Periodicity

A sequence is said to have symbol periodicity if at least one symbol is repeated periodically. Elfeky et al. [12] detect symbol periodicity in their research work. A pattern consisting of more than one symbol repeating with same periodicity in a sequence leads to sequence periodicity. If the whole sequence can be mostly represented as a repetition of a pattern or segment then that type of periodicity is called segment or segment-cycle periodicity. Sheng et al. [15] detect sequence periodicity. Elfeky et al. [12] detect full-cycle periodicity.

4.2.5 Dense Periodic Patterns

A dense fragment is a segment of time series where the distance between every pair of consecutive occurrences of the symbol is less than the distance threshold. Sheng et al. [15] detect dense periodic patterns.

There are several algorithms that discover the frequent periodic patterns [11] having minimum number of repetitions or with minimum confidence (ratio between number of occurrences found and maximum possible occurrences) eg. [12], [13]. Chitharanjan et al. [13] gives a survey on periodicity detection algorithms in time-series databases which includes Time Wraping Periodic Detection (WARP), Parial Periodic Detection Algorithm (Parper), Periodic Detection using convolusion method (CONV), Periodic Detection using SuffixTree (STNR) and also their comparison. Time-series analysis as per prior study can be divided into two types of algorithms: the first category algorithms ask the user to specify the period (or the maximum period) and then look only for patterns occurring with the specified period (or up to the maximum period), and the second class are algorithms which look for all possible periods in the time series. Yang et al. [5] mine partial periodic patterns and require the user to provide the expected period value, and then check the time series for the patterns that are periodic with that period value. A user might check the power consumption of his building for weekly, biweekly, or monthly periods. It is usually difficult to provide expected period value; it is possible that the given period value will misguide the algorithm in the detection of interesting patterns. Sheng et al. [15] presented their algorithm to find dense periodic patterns. But it required the user to provide the maximum period value. The maximum period value is difficult to be defined by the user. If the user defines; some interesting periodic patterns might be lost. The maximum period value is more difficult to be provided while discovering outlier patterns as these types of patterns may have large period value. The economic turndown pattern will have unusual periodicity; may be in years [2]. Guo et al. [10] a new class of X-outliers that have abnormal power consumption levels related to periodicity (X-axis) and propose a novel solution to detect these outliers. But the periodicity was assumed for outlier detection. A trend based periodicity detection algorithm is used for time series data with unknown periodicity.

4.3 Algorithms to detect periodic outlier patterns

Periodicity detection of outlier patterns is an area of research which is yet to be explored. It is important to note that surprising, unusual, or outlier patterns are different from outlier (values) in the data. The definition of periodic outlier detection is already mentioned in the section Detecting Periodic Outlier Pattern (section 2.4).

Rasheed et al. [2] proposed STNR-out (Suffix Tree based Noise Resilient-outlier detection) to detect periodic outlier patterns. This algorithm attempts to find three types of periodicity, namely symbol, sequence (or partial), and segment (or full-cycle) periodicity. The algorithm constructs a suffix tree for the input sequence and the tree is annotated such that each internal node records the length of the substring it represents and the frequency of the substring in the sequence. This is followed by the building of the Pattern frequency table (PFT) which will record the frequency of patterns of different length. The Candidate outlier patterns are found next and finally STNR (Suffix Tree based noise resilient) is executed for all candidate outlier patterns to output valid periodic outlier patterns. The algorithm in detail can be found in [2] and [16]. The notion of a surprising or unusual pattern considers the relative frequency of a pattern with patterns of similar length. The coverage area of the pattern and the likelihood of pattern occurrence to classify it as an outlier pattern are also considered. It can also identify outlier patterns that may involve some (or all) frequent events, as it checks the repetitions of combination of events and not just the individual events. STNR-out outlier detection algorithm is an extension of the STNR periodicity detection framework proposed by Rasheed et al. [16]. The algorithm can also detect periodic patterns in subsections, drifted or shifted periodic occurrence using time tolerance window. It can also work with noisy series containing any of insertion, deletion, and replacement noise. Rasheed et al. [2] have tested the algorithm in Walmart's Timed Transaction data set and Stock Market data to detect periodic surprising patterns successfully. Janani et al. [17] find periodicity of outlier patterns for weather forecasting system where their proposed system used suffix tree for tree formation and outlier pattern detection algorithm for predicting the outlier pattern in the weather datasets so as to improve the accuracy and response time. Dhwani et al. [18] proposed an improved framework for periodic outlier pattern detection in time series sequences. Enhancement of the algorithm proposed in [2] with MAD (Median Absolute Deviation) in place of mean of frequencies of the original paper [2] is performed. The authors have

mentioned that use of MAD increases the output of these algorithms and gives more accurate information. The authors have used a time series dataset of monthly oil prices to detect periodic outlier patterns successfully.

5. Conclusion

Outlier patterns are unusual patterns that rarely occur, and, thus, have lesser support (frequency of appearance) in the data. There are many methods such as traditional methods, soft computing methods etc. for outlier detection. This area has been undergoing research for the past few years. There are a wide variety of models developed to capture different features for outlier pattern detection in time series data. It is not possible to declare a best algorithm for outlier detection. For instance the unsupervised anomaly detection problem, the nearest neighbor based algorithms tend to be more capable of accurately identifying outliers but perform weakly for large datasets. On the other hand, clustering based anomaly detection has theoretically a lower computational effort, such that it could be preferred in cases where large datasets have to be processed. Thus one particular method will turn out to be the best for a particular kind and distribution of the data being processed. If no information is available on the data to process or if the data features can change through time in a non predictable way, than probably the best solution is to try different methods and apply a combination of many outlier detection methods which are based on different principles. Many hybrid methods are also being explored for outlier detection.

Focusing on periodic outlier patterns, the future work possible in this area includes the following:

- Expand periodic outlier detection in fuzzy time series.
- Statistical analysis leading to data summarization which will help in forecasting of events and decision making.
- Prediction of events
- Better Recommender systems
- Complex data types such as social streams in which two different data types are present in combination (text and structure) have also not been studied.

References

[1] E. Keogh, S. Lonardi, and B. Y.-C. Chiu, “Finding surprising patterns in a time series database in linear time and space,” in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2002, pp. 550–556.

[2] Rasheed, Faraz, and Reda Alhajj. "A Framework for Periodic Outlier Pattern Detection in Time-Series Sequences." (2014): 1-14, IEEE Transactions on, 44(5):569{582, May 2014.

[3] Cheboli, Deepthi. Anomaly detection of time series. Diss. University of Minnesota, 2010.

[4] J. Yang, W. Wang, and P. S. Yu, “Infominer: Mining surprising periodic patterns,” in Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining, 2001, pp. 395–400.

[5] J.Yang, W.Wang, and P.Yu, “InfoMiner+: Mining partial periodic patterns with gap penalties,” in Proc. IEEE Int. Conf. Data Mining, Dec. 2002, pp. 725–728.

[6] Chuah, Mooi Choo, and Fen Fu. "ECG anomaly detection via time series analysis." Frontiers of High Performance Computing and Networking ISPA 2007 Workshops. Springer Berlin Heidelberg, 2007.

[7] Fu, Ada Wai-Chee, et al. "Finding time series discords based on haar transform." Advanced Data Mining and Applications. Springer Berlin Heidelberg, 2006. 31-41.

[8] Keogh, Eamonn, Jessica Lin, and Ada Fu. "Hot sax: Efficiently finding the most unusual time series subsequence." Data mining, fifth IEEE international conference on. IEEE, 2005.

[9] Lin, Yi, Michael D. McCool, and Ali A. Ghorbani. "Motif and anomaly discovery of time series based on subseries join." IAENG International Conference on Data Mining and Applications, ICDMA. 2010.

[10] Guo, Zhihui. “X-outlier detection and periodicity detection in load curve data in power systems.” Diss. Applied Science: School of Computing Science, 2011.

[11] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, “Periodicity detection in time series databases,” IEEE Trans. Knowl. Data Eng., vol. 17, no. 7, pp. 875–887, Jul. 2005.

[12] M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid, “WARP: Time warping for periodicity detection,” in Proc. IEEE Int. Conf. Data Mining, Nov. 2005, pp. 8–15.

[13] Chitharanjan, K. "PERIODICITY DETECTION ALGORITHMS IN TIME SERIES DATABASES-A SURVEY.", International Journal of Computer Science & Engineering Technology, Jan 2013

[14] Sirisha, M. Shashi, and G. N. V. G. GV Padma Raju. "Periodic Pattern Mining–Algorithms and Applications." Global Journal of Computer Science and Technology 13.13 (2014).

[15] C. Sheng, W. Hsu, and M.-L. Lee, “Mining dense periodic patterns in time series data,” in Proc. IEEE Int. Conf. Data Eng., 2005, p. 115.

[16] Rasheed, Faraz, Mohammed Alshalalfa, and Reda Alhajj. "Efficient periodicity mining in time series databases using suffix trees." Knowledge and Data Engineering, IEEE Transactions on 23.1 (2011): 79-94.

[17] Janani, B., and S. Rajkumar. "Periodicity Detection of Outlier Pattern for Weather Forecasting System.", International conference on Simulations in Computing Nexus, March 2014.

[18] Dave, Dhwani, and Tanvi Varma. "An improved framework for periodic outlier pattern detection in time series sequences.", International Journal of Emerging Trends in Engineering and Development, July 2014

[19] Gupta, Manish, et al. "Outlier Detection for Temporal Data." Synthesis Lectures on Data Mining and Knowledge Discovery 5.1 (2014): 1-129.

[20] Lahiri, Mayank, and Tanya Y. Berger-Wolf. "Mining periodic behavior in dynamic social networks." Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. IEEE, 2008.