Unsupervised Subsequence Anomaly Detection in Large Sequences

Paul Boniol
Supervised by: Themis Palpanas, Mohammed Meftah, Emmanuel Remy
EDF R&D, LIPADE, Université de Paris
paul.boniol@edf.fr

ABSTRACT

Subsequence anomaly detection in long sequences is an important problem with applications in a wide range of domains. However, the approaches that have been proposed so far in the literature have severe limitations: they either require prior domain knowledge that is used to design the anomaly discovery algorithms, or become cumbersome and expensive to use in situations with recurrent anomalies of the same type. In this Ph.D. work, we address these problems, and present unsupervised methods suitable for domain agnostic subsequence anomaly detection. We explore two possible way to represent the normal behavior of a long data series that lead to fast and accurate identification of abnormal subsequences. These normal representations are either based on subsequences (using a data structure called the normal model), or on graphs, by taking advantage of graph properties to encode the normal transitions between neighboring subsequences of a long series. The experimental results, on a large set of synthetic and real datasets, demonstrate that the proposed approaches correctly identify single and recurrent anomalies of various types, without any prior knowledge of the characteristics of these anomalies. Our approaches outperform by a large margin several competing approaches in accuracy, while being up to orders of magnitude faster.

1. INTRODUCTION

Time series[1] anomaly detection is a crucial problem with application in a wide range of domains. Examples of such applications can be found in manufacturing, astronomy, engineering, and other domains [10, 11]. This implies a real need by relevant applications for developing methods that can accurately and efficiently achieve this goal.

Anomaly detection is a well studied task [2, 13, 16, 8] that can be tackled by either examining single values, or sequences of points. In the specific context of sequences, which is the focus of this paper, we are interested in identifying anomalous subsequences [16, 12, 8], which are not single abnormal values, but rather an abnormal *sequence* of values. In real-world applications, this distinction becomes crucial: in certain cases, even though every individual point may be normal, the trend exhibited by the sequence of these same values may be anomalous. Evidently, failing to identify such situations could lead to severe problems that are only detected when it is too late.

Some existing techniques explicitly look for a set of predetermined types of anomalies [1]. These are techniques that have been specifically designed to operate in a particular setting, they require domain expertise, and cannot generalize. Other techniques identify as anomalies the subsequences with the largest distances to their nearest neighbors (termed discords) [16, 12]. The assumption is that the most distant subsequence is completely isolated from the "normal" subsequences.

However, this definition fails in the case where an anomaly repeats itself (approximately the same). In this situation, anomalies will have other anomalies as close neighbors, and will not be identified as discords. In order to remedy this situation, the $m^{th} discord$ approach has been proposed [15], which takes into account the multiplicity m of the anomalous subsequences that are similar to one another, and marks as anomalies all the subsequences in the same group. However, this approach assumes that we know the cardinality of the anomalies, which is not true in practice (otherwise, we need to try several different m values, increasing drastically the execution time). Furthermore, the majority of the previous approaches require prior knowledge of the anomaly length, and their performance deteriorates significantly when the correct length value is not used.

Our work addresses the aforementioned problems. We proposed two approaches aiming to build a normal representation of the data series, which enables to identify anomalous subsequences. The two approaches use different data structures to represent the normal behavior of the sequences; they are summarized below:

- NormA [3, 4], a normal-model based subsequence anomaly detection method in large data series, for which the normal data structure is a set of subsequences paired with a weight. The higher those weights are, the more "normal" their paired subsequences are.

[1] A time series, or data series, or sequence, is an ordered sequence of real-valued points. If the dimension that imposes the ordering is time then we talk about *time series*, but it could also be mass, angle, position, etc. We will use the terms *time series*, *data series*, and *sequence* interchangeably.

Proceedings of the VLDB 2020 PhD Workshop, August 31st, 2020. Tokyo, Japan.

- Series2Graph [5, 6], an unsupervised graph-based approach for subsequence anomaly detection in large data series, for which the data structure is modeled as a directed graph. Nodes fo this graph can be seen as subsequences of the data series.

2. NORMAL BEHAVIOR

In this section we describe our proposed approaches to the aforementioned problems. In the next part we describe the notions and the elements used in our solutions.

We begin by introducing some formal notations useful for the rest of the paper. A data series $T \in \mathbb{R}^n$ is a sequence of real-valued numbers $T_i \in \mathbb{R}$ $[T_1, T_2, ..., T_n]$, where $n = |T|$ is the length of T, and T_i is the i^{th} point of T. We are typically interested in local regions of the data series, known as subsequences. A subsequence $T_{i,\ell} \in \mathbb{R}^\ell$ of a data series T is a continuous subset of the values from T of length ℓ starting at position i. Formally, $T_{i,\ell} = [T_i, T_{i+1}, ..., T_{i+\ell-1}]$. Given two sequences, A and B, of the same length, ℓ, we can calculate their Z-normalized Euclidean distance, *dist*, as follows: $dist = \sqrt{\sum_{i=1}^{\ell}(\frac{A_i - \mu_A}{\sigma_A} - \frac{B_i - \mu_B}{\sigma_B})^2}$, where μ and σ represent respectively the mean and standard deviation of the sequences. For the remaining of this paper, we will simply use the term *distance.*

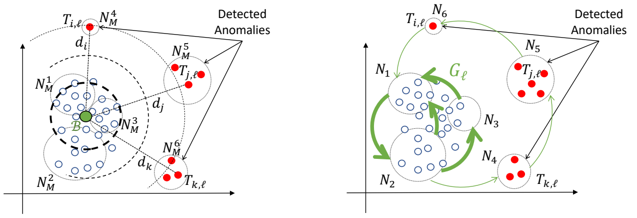
Given a subsequence $T_{i,\ell}$, we say that its m^{th} Nearest Neighbor (m^{th} NN) is $T_{j,\ell}$ if $T_{j,\ell}$ has the m^{th} shortest distance to $T_{i,\ell}$ among all the subsequences of length ℓ in T, excluding trivial matches; a trivial match of $T_{i,\ell}$ is a subsequence $T_{a,\ell}$, where $|i - a| < \ell/2$ (i.e., the two subsequences overlap by more than half their length). Since we are interested in subsequence anomalies, we first define the set of all subsequences of length ℓ in a given data series T: $\mathbb{T}_\ell = \{T_{i,\ell} | \forall i. 0 \le i \le |T| - \ell + 1\}$. In general, we assume that $\mathbb{T}_\ell$ contains both normal and anomalous subsequences. We then need a way to characterize normal behavior. For the purpose of the paper, we abstractly define the normal behavior of the data series as follows:

DEFINITION 1 (NORMAL BEHAVIOR, N_B). *Given a data series T, N_B is a model that represents the normal (i.e., not anomalous) trends and patterns in T.*

The above definition is not precise on purpose: it allows several interpretations, which can lead to different kinds of models. Nevertheless, subsequence anomalies can then be defined in a uniform way: anomalies are the subsequences that have the largest distances to the expected, normal behavior, N_B (or their distance is above a set threshold). Both of the proposed approaches will define their own data structure to represent N_B, as well as their own distance measure.

2.1 The NormA Approach

We describe a new approach for subsequence anomaly detection. We propose a formalization for N_B, called the Normal Model, denoted N_M, and defined as follows [3]: N_M is a set of sequences, $N_M = \{(N_M^0, w^0), (N_M^1, w^1), ..., (N_M^n, w^n)\}$, where N_M^i is a subsequence of length ℓ_{N_M} (the same for all N_M^i) that corresponds to a recurring behavior in the data series T, and w^i is its normality score (as we explain later, the highest this score is, the more usual the behavior represented by N_M^i is). In other words, this model averages (with proper weights) the different recurrent behaviors observed in the data, such that all the normal behaviors of the data series will be represented in the normal model, while unusual behaviors will not (or will have a very low weight).



(a) *d* defined as the distance to the Normal Model $N_M = \{(N_M^1, w^1), \dots, (N_M^6, w^6)\}$; $\mathcal{B}$ is the weighted barycenter of $\mathrm{N_M}$

(b) Normal represtnation of the data series as G_ℓ .The path to reach the red dots has smaller weight than those to reach the blue dots.

Figure 1: Illustration of the two subsequence anomaly definitions proposed approaches: (a) NormA; (b) Series2Graph.

Figure 1(a) is an illustration of a Normal Model. As depicted, the Normal Model N_M is a weighted combination of a set of subsequences (points within the dotted circles). The combination of these subsequences and their related weights returns distances d_i, d_j, d_k that are high enough to be differentiated from the normal points/subsequences. These distances can be seen as the distance between subsequences and a weighted barycenter $\mathcal{B}$ (in green) that represents N_M. Note that we do not actually compute this barycenter; we illustrate it in Figure 1(a) for visualization purposes.

We choose $\ell_{N_M} > \ell$ in order to make sure that we do not miss useful subsequences, i.e., subsequences with a large overlap with an anomalous subsequence. For instance, for a given subsequence of length ℓ, a normal model of length $\ell_{N_M} = 2\ell$ will also contain the subsequences overlapping with the first and last half of the anomalous subsequence. We can now define the Abnormal subsequences as follows [3].

DEFINITION 2 (SUBSEQUENCE ANOMALY). *Assume a data series T, the set $\mathbb{T}_\ell$ of all its subsequences of length ℓ, and the Normal Model N_M of T. Then, the subsequence $T_{j,\ell} \in \mathbb{T}_\ell$ with* anomaly score, *i.e., distance to N_M, $d_j = \sum_{N_M^i} w^i * min_{x \in [0, \ell_{N_M} - \ell]}\{dist(T_{j,\ell}, (N_M^i)_{x,\ell})\}$ is an anomaly if d is in the Top-k largest distances among all subsequences in $\mathbb{T}_\ell$, or $d > \epsilon$, where $\epsilon \in \mathbb{R}_{>0}$ is a threshold.*

Note that the only essential input parameter is the length ℓ of the anomaly (which is also one of the inputs in all relevant algorithms in the literature [12, 16, 7, 9, 8]). The parameter k (or ϵ) is not essential, as long as the algorithm can *rank* the anomalies.

We stress that in practice, experts start by examining the most anomalous pattern, and then move down in the ranked list, since there is (in general) no rigid threshold separating anomalous from non-anomalous behavior [2].

As we mentioned above and will detail later on, we choose to define N_M as a set of sequences that summarizes normality in T, by representing the average behavior of a set of normal sequences. Intuitively, N_M is the set of data series, which tries to minimize the sum of Z-normalized Euclidean distances between itself and some of the subsequences in T. Last but not least, we need to compute N_M in an unsupervised way, i.e., without having normal/abnormal labels for the subsequences in $\mathbb{T}_\ell$.

Observe that this definition of N_M implies the following challenge: even though N_M summarizes the normal behavior only, it needs to be computed based on T, which may include (several) anomalies. We address these challenges by taking advantage of the fact that anomalies are a minority class.

2.1.1 NormA Framework

We now briefly describe NormA [3], our unsupervised solution to the subsequence anomaly detection problem.

[**Computing the Normal Model**] Recall that N_M should capture (summarize) the normal behavior of the data. This may not be very hard to do for a sequence T that does *not* contain any anomalous subsequences. In practice though, we want to apply the NormA approach in an unsupervised way on any sequence, which may contain several anomalies.

We compute the N_M in three steps. First, we extract the subsequences that can serve as candidates for building the N_M. These candidates are either randomly selected from T (NormA-smpl), or correspond to motifs. Then, we group these subsequences according to their similarity in a set of clusters $\mathbb{C}$ (we use hierarchical clustering and Minimum Description Length to identify the right number of clusters). The last step consists of scoring each cluster, and selecting the cluster that best represents normal behavior. Formally, for a given cluster $c \in \mathbb{C}$, we set a weight $w^c = Norm(c, \mathbb{C})$ with the following formula: $Norm(c, \mathbb{C}) = \frac{Frequency(c)^2 \times Coverage(c)}{\sum_{x \in \mathbb{C}} dist(Center(c), Center(x))}$, where $Frequency(c)$ is the number of subsequences in c, and $Coverage(c)$ is the time interval between the first and the last occurrence of a subsequence in c. We thus set $N_M = \{(N_M^0, w^0), (N_M^1, w^1), ..., (N_M^n, w^n)\}$, with N_M^i the centroid of the i^{th} cluster in $\mathbb{C}$.

[**Normal Model Based Anomaly Detection**] Intuitively, the anomalous subsequences of the long series T are the ones that are far away from N_M. We thus compute the meta $S = [d_0, d_1, ..., d_{|T|-\ell_{N_M}}]$, with d_j the distance of the subsequence $T_{i,j}$ to the Normal Model N_M as defined in Definition 2. These distances correspond to the degree of abnormality: the larger the distance is to N_M, the more abnormal the subsequence is. We then extract the k subsequences of length ℓ with the highest distances to N_M, and rank them according to their distances. Alternatively, we can extract all subsequences with distance larger than a threshold.

2.2 The Series2Graph Approach

We now present an alternative data structure to represent the subsequences normal behaviors of the data series. The previous normal model was a set of subsequences that aimed to store both normal and abnormal subsequences into the same set, associated with weights that rank them based on their normality. One can argue that an ordering information is missing from this data structure representation. We thus formulate an approach for subsequence anomaly detection based on the data series representation into a Graph, in which edges encode the ordering information [5]. Figure 1(b) illustrates the Graph data structure.

We first define several basic elements related to graphs. We define a Node Set $\mathcal{N}$ as a set of unique integers. Given a Node Set $\mathcal{N}$, an Edge Set $\mathcal{E}$ is then a set composed of tuples (x_i, x_j), where $x_i, x_j \in \mathcal{N}$. $w(x_i, x_j)$ is the weight of that edge. Given a Node Set $\mathcal{N}$, an Edge Set $\mathcal{E}$ (pairs of nodes in $\mathcal{N}$), a Graph G is an ordered pair $G = (\mathcal{N}, \mathcal{E})$. A directed graph or digraph G is an ordered pair $G = (\mathcal{N}, \mathcal{E})$ where $\mathcal{N}$ is a Node Set, and $\mathcal{E}$ is an ordered Edge Set.

We now provide a new formulation for subsequence anomaly detection. The idea is that a data series is transformed into a sequence of abstract states (corresponding to different subsequence patterns), represented by nodes $\mathcal{N}$ in a directed graph, $G(\mathcal{N}, \mathcal{E})$, where the edges $\mathcal{E}$ encode the number of times one state occurred after another. Under this formulation, paths in the graph composed of high-weight edges and high-degree nodes correspond to normal behavior. Then, the Normality of a data series is defined as follows [5].

DEFINITION 3 (θ-NORMALITY). *Let a node set be defined as $\mathcal{N} = \{N_1, N_2, ..., N_m\}$. Let also a data series T be represented as a sequence of nodes $\langle N^{(1)}, N^{(2)}, ..., N^{(n)} \rangle$ with $\forall i \in [0, n], N^{(i)} \in \mathcal{N}$ and $m \leq n$. The θ-Normality of T is the subgraph $G_\theta^\nu(\mathcal{N}_\nu, \mathcal{E}_\nu)$ of $G(\mathcal{N}, \mathcal{E})$ with $\mathcal{E} = \{(N^{(i)}, N^{(i+1)})\}_{i \in [0, n-1]}$, such that: $\mathcal{N}_\nu \subset \mathcal{N}$ and $\forall (N^{(i)}, N^{(i+1)}) \in \mathcal{E}_\nu, w((N^{(i)}, N^{(i+1)})).(deg(N^{(i)}) - 1) \geq \theta$.*

Using θ-Normality subgraphs naturally leads to a ranking of subsequences based on their "normality". For practical reasons, this ranking can be transformed into a score, where each rank can be seen as a threshold in that score. We used such a score in GraphAn to detect abnormal subsequences.

Note that given the existence of graph G, the above definitions imply a way for identifying the anomalous subsequences. The problem is now how to construct this graph.

2.2.1 Series2Graph Framework

We now briefly describe Series2Graph [5, 6], our unsupervised solution to the subsequence anomaly detection problem. For a given data series T, the overall Series2Graph process is divided into four main steps as follows:

[**Subsequence Embedding**] We project all the subsequences (of a given length ℓ) of T in a three-dimensional space that corresponds to the three most important components of the Principle Component Analysis (PCA). This space is subsequently transformed into a two-dimensional space, composed of two components $\vec{r_y}, \vec{r_z}$ corresponding to two orthogonal vectors of $\vec{v_{ref}}$, an axis composed of every flat sequences ($a * \mathbf{1}_{\ell-\lambda}$ with $a \in \mathbb{R}$). In the later space, the shape similarity is preserved [5].

[**Node Creation**] We then create a node for each one of the densest parts of the above two-dimensional space. The space is first discretized by a set of radius $\mathcal{I}_\psi$ of angle ψ. We then estimate the density of subsequences along each one of these radius using gaussian kernels. The maximal values of the estimated densities are assigned to nodes and form our Node Set $\mathcal{N}$. These nodes summarize all major patterns of length ℓ that occurred in T [5].

[**Edge Creation**] We then retrieve all transitions between pairs of subsequences represented by two different nodes: each transition corresponds to a pair of subsequences, where one occurs immediately after the other in the input data series T. We represent transitions with an edge between the corresponding nodes, and we thus form the Edge Set $\mathcal{E}$. The edge weights are set to the number of times the corresponding pair of subsequences was observed in T. Finally we build our graph $G_\ell(\mathcal{N}, \mathcal{E})$.

[**Subsequence Scoring**] We compute the normality (or anomaly) score of a subsequence of length $\ell_q \geq \ell$

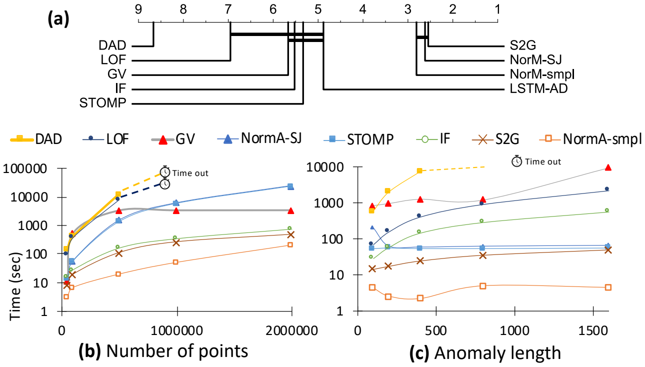


Figure 2: Comparison of six state-of-the-art methods to Series2graph and NormA: (a) Precision@k with a critical diagram over 21 datasets; (b) execution time vs varying number of points; (c) execution time vs varying anomaly length.

(within or outside of T), based on the previously computed edges/nodes and their weights/degrees. Formally, for a subsequence T_{i,ℓ_q} of T, represented by a path in $G_\ell(\mathcal{N},\mathcal{E})$ $P_{th} = \langle N^{(i)}, N^{(i+1)}, ..., N^{(i+\ell_q)} \rangle$, the normality score is defined as: $Norm(P_{th}) = \sum_{j=i}^{i+\ell_q-1} \frac{w(N^{(j)},N^{(j+1)})(deg(N^{(j)})-1)}{\ell_q}$, where $w(e)$ and $deg(n)$ are the weight of edge e and the degree of node n, respectively.

2.3 Experimental Analysis

In our preliminary experimental evaluations [3, 5], we used 21 synthetic and real datasets from diverse domains, and measured Precision@k (i.e., accuracy in the Top-k answers) and execution time. (Due to lack of space, we only report here aggregated results.)

After rejecting the null hypothesis using the Friedman test, we used the pairwise Post-Hoc Analysis with a Wilcoxon signed-rank test ($\alpha = 0.05$) [14], and we depict the results in Figure 2(a). The results show that the NormA and Series2Graph approaches are the overall winners, performing statistically significantly better than the state-of-the-art algorithms from both the data series and the multidimensional data communities. (Even though Series2Graph seems to be slightly better than both variants of Norma, this difference is not statistically significant.)

Moreover, Figures 2(b) and (c) demonstrate that NormA-smpl and Series2Graph are considerably faster than the other methods (note the log-scale y-axis). In particular, NormA-smpl is between 1-3 orders of magnitude faster than the rest, as we increase the length of the input sequence, or the length of the anomalies.

We note that further experiments are needed in order to compare in detail our proposed methods.

3. CONCLUSIONS AND FUTURE WORK

Our studies on NormA and Series2Graph, described above, constitute a first attempt to formalize the problem of Normal Behavior representation in the context of data series. The two proposed approaches describe data structures based on subsequence sets and graphs, respectively. The experimental results show that both NormA and Series2Graph outperform the current state-of-the-art techniques methods over a large set of diverse datasets. This indicates that the proposed approach of using a Normal Behavior model for the subsequence anomaly detection problem is very promising.

In our current work, we focus on the following directions. [**Normal Behavior Model**] We will study in depth different models for Normal Behavior, and perform analytical and experimental comparisons among them. The goal will be to understand the theoretical properties, as well as the practical limitations of each one, so as to make informed decisions about the best model to use.
[**Multivariate Data Series**] Given that many domains monitor, process, and analyze multivariate data series, we will work on algorithms that will inherently consider together all variables of the series. The challenging problems are how to create a multivariate Normal Behavior model and corresponding data structure, and how to define and formalize suitable distance measures for this case.
[**Online Operation**] In several situations, subsequence anomaly detection applications need to operate in an online fashion. In this context, we will develop extensions of our methods for real-time applications. We will analyze the impact of data characteristics changes on the Normal Behavior models, and explore solutions for updating the Normal Behavior data structures in real-time.

Acknowledgments Work partially supported by EDF R&D and ANRT French program.

References

[1] D. Abboud, M. Elbadaoui, W. Smith, and R. Randall. Advanced bearing diagnostics: A comparative study of two powerful approaches. *MSSP*, 114, 2019.

[2] V. Barnet and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons, Inc., 1994.

[3] P. Boniol, M. Linardi, F. Roncallo, and T. Palpanas. Automated Anomaly Detection in Large Sequences. In *ICDE*, 2020.

[4] P. Boniol, M. Linardi, F. Roncallo, and T. Palpanas. SAD: An Unsupervised System for Subsequence Anomaly Detection. In *ICDE*, 2020.

[5] P. Boniol and T. Palpanas. Series2graph: Graph-based subsequence anomaly detection for time series. *PVLDB*, 13(11), 2020.

[6] P. Boniol, T. Palpanas, M. Meftah, and E. Remy. Graphan: Graph-based subsequence anomaly detection. *PVLDB*, 13(11), 2020.

[7] Y. Bu, O. T. Leung, A. W. Fu, E. J. Keogh, J. Pei, and S. Meshkin. WAT: finding top-k discords in time series database. In *SIAM*, 2007.

[8] M. Linardi, Y. Zhu, T. Palpanas, and E. J. Keogh. Matrix Profile Goes MAD: Variable-Length Motif And Discord Discovery in Data Series. In *DAMI*, 2020.

[9] W. Luo and M. Gallagher. Faster and parameter-free discord search in quasi-periodic time series. In *Advances in Knowledge Discovery and Data Mining*, 2011.

[10] T. Palpanas. Data series management: The road to big sequence analytics. *SIGMOD Rec.*, 44(2):47–52, Aug. 2015.

[11] T. Palpanas and V. Beckmann. Report on the First and Second Interdisciplinary Time Series Analysis Workshop (ITISA). *ACM SIGMOD Record*, 48(3), 2019.

[12] P. Senin, J. Lin, X. Wang, T. Oates, S. Gandhi, A. P. Boedihardjo, C. Chen, and S. Frankenstein. Time series anomaly discovery with grammar-based compression. In *EDBT*, 2015.

[13] S. Subramaniam, T. Palpanas, D. Papadopoulos, V. Kalogeraki, and D. Gunopulos. Online outlier detection in sensor data using non-parametric models. In *VLDB 2006*, pages 187–198, 2006.

[14] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[15] D. Yankov, E. J. Keogh, and U. Rebbapragada. Disk aware discord discovery: finding unusual time series in terabyte sized datasets. *Knowl. Inf. Syst.*, 17(2):241–262, 2008.

[16] C. M. Yeh, Y. Zhu, L. Ulanova, N. Begum, Y. Ding, H. A. Dau, D. F. Silva, A. Mueen, and E. J. Keogh. Matrix profile I: all pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In *ICDM*, pages 1317–1322, 2016.