SCIENCE CHINA
Information Sciences

• LETTER •

April 2021, Vol. 64 149101:1–149101:3
https://doi.org/10.1007/s11432-020-3021-6

SPAR: set-based piecewise aggregate representation for time series anomaly detection

Peng ZHAN[1,3], Yupeng HU[2*], Lin CHEN[1,3], Wei LUO[1] & Xueqing LI[1]

[1]*School of Software, Shandong University, Jinan* 250100, *China;*
[2]*School of Computer Science and Technology, Shandong University, Qingdao* 266237, *China;*
[3]*Informatization Office, Shandong University, Jinan* 250100, *China*

Received 1 January 2020/Revised 13 April 2020/Accepted 4 June 2020/Published online 26 February 2021

Citation Zhan P, Hu Y P, Chen L, et al. SPAR: set-based piecewise aggregate representation for time series anomaly detection. Sci China Inf Sci, 2021, 64(4): 149101, https://doi.org/10.1007/s11432-020-3021-6

Dear editor,

Time series anomaly detection, aiming for identifying unexpected observations within the given time series, has been considered as one of the most challenging studies in time series data mining [1, 2]. In this study, we present a novel set-based piecewise aggregate representation (SPAR) for anomaly detection, dubbed as SPAR-AD. Compared with other existing anomaly sequence detection methods, SPAR-AD not only pays close attention to recognize the significant changes of time series in the amplitude domain, but also keeps a watchful eye on identifying the corresponding variations in the temporal domain. Concretely, SPAR-AD can evenly divide a given time series into non-overlapping sequences and further calculate the corresponding anomaly scores of these sequences based on their own amplitude temporal features. Accordingly, all anomaly sequences with relatively high anomaly scores in the given time series can be detected effectively.

Set-based piecewise aggregate representation. For a given time series T, expressed as $T = \{v_1, v_2, \ldots, v_i, \ldots, v_n\}$, where the values in T are strictly ordered by their own time stamps. Considering that we are concerned with detecting anomalies in the local sequences rather than the entire T, traditional sliding window (SW) approaches [2, 3] with length m can be utilized for dividing T into the sequence set (SS) containing $K(\lceil n/m \rceil)$ sequences, expressed as $\mathrm{SS} = \{S_1, S_2, \ldots, S_k, \ldots, S_K\}$. The k-th sequence in SS, can be indicated as $S_k = \{v_{k\times m}, v_{k\times m+1}, \ldots, v_{k\times m+m-1}\}$. Thereafter, all the sequences are subdivided into non-overlapping subsequences, in accordance with their temporal order, and further projected into the corresponding subregions of the multi-domain space. Without loss of generality, supposing the k-th sequence S_k can be subdivided into $p = m/L$ (m can be evenly divided by L) subsequences, the k-th subsequence set $\mathrm{Sub}S_k$ is expressed as $\mathrm{Sub}S_k = \{\mathrm{sub}_1^k, \ldots, \mathrm{sub}_j^k, \ldots, \mathrm{sub}_p^k\}$, where each sub_j^k is denoted as $\mathrm{sub}_j^k = \{v_{k\times j}, v_{k\times j+1}, \ldots, v_{k\times j+L-1}\}$.

Subsequently, motivated by the novel multi-resolution time series representation [4], each $\mathrm{Sub}S_k$ is projected into the specific temporal and amplitude domain space MS_k, including $p \times q$ equal-area regions, expressed as

$$\mathrm{MS}_k = \begin{bmatrix} R_{q,1}^k & R_{q,2}^k & \cdots & R_{q,c}^k & \cdots & R_{q,p}^k \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ R_{t,1}^k & R_{t,2}^k & \cdots & R_{t,c}^k & \cdots & R_{t,p}^k \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ R_{1,1}^k & R_{1,2}^k & \cdots & R_{1,c}^k & \cdots & R_{1,p}^k \end{bmatrix}, \tag{1}$$

where p, q are denoted as the number of equal divisions in temporal and amplitude domain, respectively. $R_{t,c}^k$ is further expressed as $R_{t,c}^k = \{[k \times c : k \times (c+1) - 1], [v_{\min} : v_{\max}]\}$. $[k \times c : k \times (c+1) - 1]$ is the temporal range of $R_{t,c}^k$, $v_{\min}$ and $v_{\max}$ are the lower and upper bounds of $R_{t,c}^k$ in the amplitude domain.

With the help of multi-domain space MS_k, $\mathrm{Sub}S_k$ is transformed into the corresponding representation result SR_k in accordance with the constraints of the amplitude and duration on subregions, expressed as

$$\mathrm{SR}_k = \begin{bmatrix} E_{q,1}^k & E_{q,2}^k & \cdots & E_{q,c}^k & \cdots & E_{q,p}^k \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ E_{t,1}^k & E_{t,2}^k & \cdots & E_{t,c}^k & \cdots & E_{t,p}^k \\ \vdots & \cdots & \cdots & \cdots & \cdots & \vdots \\ E_{1,1}^k & E_{1,2}^k & \cdots & E_{1,c}^k & \cdots & E_{1,p}^k \end{bmatrix}, \tag{2}$$

where $E_{t,c}^k$ is the set-based piecewise aggregate approximation (PAA) representation result on $R_{t,c}^k$. Concretely, the mean value ($\mathrm{mv}_{t,c}^k$) and the duration ratio $\omega_{t,c}^k$ ($\omega_{t,c}^k = d_t/m$) are used to represent the amplitude, temporal features of the specific subsequence in $R_{t,c}^k$. Therefore, the entire representation results SRS on SS are expressed as $\mathrm{SRS} = \{\mathrm{SR}_1, \mathrm{SR}_2, \ldots, \mathrm{SR}_k, \ldots, \mathrm{SR}_K\}$.

* Corresponding author (email: huyupeng@sdu.edu.cn)

SPAR for anomaly detection. With the help of SPAR, all the sequences in SS can be represented by standing on a multi-domain view. Supposing the corresponding representation results of two sequences (S_k, S_h) are SR_k and SR_h, we propose a novel anomaly evaluation strategy to calculate the corresponding anomaly scores of S_k and S_h. The difference between SR_k and SR_h can be calculated as follows:

$$\begin{aligned} D(\mathrm{SR}_k, \mathrm{SR}_h) &= \|\mathrm{SR}_k - \mathrm{SR}_h\| \\ &= \sqrt{\sum_{t=1}^{q}\sum_{c=1}^{p}(|mv_{t,c}^k - mv_{t,c}^h|^2 \times |\omega_{t,c}^k - \omega_{t,c}^h|)}. \end{aligned} \tag{3}$$

$D(\mathrm{SR}_k, \mathrm{SR}_h)$ not only considers the amplitude differences between two sequences, but also focuses on the discrepant temporal duration in different regions. Therefore, the difference between SR_k and SR_h can be evaluated more comprehensively. Moreover, the anomaly score ζ_k of S_k can be calculated as follows:

$$\zeta_k = \frac{[\sum_{j=1}^{K} D(\mathrm{SR}_k, \mathrm{SR}_j) + \sum_{j=1}^{K} D(\mathrm{SR}_j, \mathrm{SR}_k)] \times K}{\sum_{m=1}^{K}[\sum_{i=1}^{K} D(\mathrm{SR}_i, \mathrm{SR}_m) + \sum_{j=1}^{K} D(\mathrm{SR}_m, \mathrm{SR}_j)]}. \tag{4}$$

Finally, according to 4, the corresponding anomaly sequences in SS can be detected effectively.

Experiment and analysis. To evaluate the performance of our proposed SPAR-AD objectively, 23 real word time series datasets, including 20 open source datasets from UCR Time Series Archive[1)] and 3 our collected network traffic flow time series datasets of Shandong University, are selected for comparison experiments. Moreover, to thoroughly measure our method and the baselines, average anomaly confidence index, denoted as AACI, anomaly detection rate (ADR) are selected as the evaluation metrics. AACI is calculated as follows:

$$\mathrm{AACI} = \mathrm{Mean}(\mathrm{AS}), \tag{5}$$

where AS denotes the anomaly sequence set of SS (AS $\subset$ SS) and AACI is the mean anomaly score of AS. Based on a fixed threshold γ, the higher AACI is, the stronger detection ability of corresponding method has. ADR refers to the percentage of anomalies detected in the total number of anomalies, calculated as follows:

$$\mathrm{ADR} = \frac{\mathrm{PA}}{\mathrm{Num}}, \tag{6}$$

where PA refers to the number of anomalies that are detected successfully, Num refers to the total number of anomaly sequences in AS.

Besides, we chose 3 highly cited anomaly detection methods SAX-AD [5], PAA-AD [6] and APAA-AD [7] as the baseline methods for the following comparison experiments.

The corresponding comparison results of 4 methods on 23 datasets are shown in Table 1, where m denotes the length of sliding window and Num refers to the total number of anomaly sequences.

Table 1 Comparison experiments on 23 real world datasets

Data set	m	Num	SAX-AD		PAA-AD		APAA-AD		SPAR-AD	
			ADR	AACI	ADR	AACI	ADR	AACI	ADR	AACI
BeetleFly	512	2	0.00	0.00	0.00	0.00	1.00	2.13	1.00	2.24
BirdChicken	512	2	0.00	0.00	0.00	0.00	1.00	2.38	1.00	2.42
BME	128	5	1.00	2.65	1.00	3.04	1.00	2.91	1.00	4.15
ChlorineConcentration	166	7	0.71	2.58	0.71	3.40	0.85	4.54	1.00	6.61
Coffee	286	1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.27
Computers	720	0	0.00	0.00	0.00	0.00	0.00	2.17	0.00	0.00
CricketZ	300	2	0.00	0.00	0.00	0.00	0.50	2.62	1.00	2.67
Crop	46	3	0.33	2.17	0.67	2.20	1.00	2.39	1.00	3.21
DiatomSizeReduction	345	3	1.00	3.81	1.00	3.90	1.00	2.98	1.00	2.79
DistalPhalanxTW	80	7	1.00	2.48	0.85	2.49	1.00	2.76	1.00	3.99
DodgerLoopDay	288	1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.04
Earthquakes	512	1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.47
ECG200	96	1	0.00	0.00	1.00	2.10	1.00	2.42	1.00	5.05
ECG5000	140	2	0.00	0.00	1.00	2.42	1.00	2.47	1.00	4.03
Fish	463	4	0.25	2.12	0.25	2.44	0.50	2.27	0.75	2.51
GunPoint	150	3	0.00	0.00	0.33	2.23	0.33	2.12	1.00	2.66
ItalyPowerDemand	24	1	1.00	2.22	1.00	2.39	1.00	3.68	1.00	4.44
Meat	448	3	0.67	2.43	0.67	2.33	0.67	2.19	1.00	2.67
OSULeaf	427	1	0.00	0.00	0.00	0.00	0.00	0.00	1.00	2.19
Strawberry	235	2	1.00	2.38	1.00	2.54	1.00	2.76	1.00	4.03
SDUInflowNetwork	288	2	0.00	0.00	0.00	0.00	0.50	2.11	1.00	2.97
SDUOutflowNetwork	288	2	0.50	2.57	1.00	2.33	1.00	2.59	1.00	3.53
SDUTotalNetwork	288	2	0.50	2.03	0.50	2.14	0.50	2.26	1.00	2.65
Average	–	–	51%	2.49	62%	2.57	80%	2.62	98%	3.25

1) https://www.cs.ucr.edu/ eamonn/time series data 2018/.

As shown in Table 1, we have the following observations. Firstly, both SAX-AD and PAA-AD deliver inferior performance on ADR and AACI, because they represent time series sequences by either mean-value transformed symbols or roughly mean values, and thus fail to effectively recognize the corresponding anomaly sequences within the given time series. Secondly, APAA-AD has relatively higher ADR and AACI than other 2 baselines, which verifies that the amplitude domain division based representation is effective for anomaly detection. Finally, SPAR-AD achieves the highest ADR and AACI, substantially surpassing 3 baselines. Especially, the average ADR of SPAR-AD is 18% higher than that of APAA-AD, which means that based on more comprehensive multi-domain representation, SPAR-AD has stronger ability on recognizing abnormal sequences.

Acknowledgements This work was supported by National Key Research Program of China (Grant No. U1936203), Shandong Provincial Natural Science and Foundation (Grant No. ZR2019JQ23), CERNET Innovation Project (Grant No. NGII20190109), and Project of Qingdao Postdoctoral Applied Research.

References

1 Zhang Q, Hu Y P, Ji C, et al. Edge computing application: real-time anomaly detection algorithm for sensing data. J Comput Res Dev, 2018, 55: 524–536
2 Hu Y, Zhan P, Xu Y, et al. Temporal representation learning for time series classification. Neural Comput Appl, 2020, 32: 1–14
3 Hu Y, Ji C, Zhang Q, et al. A novel multi-resolution representation for time series sensor data analysis. Soft Comput, 2020, 24: 10535–10560
4 Zhan P, Sun C, Hu Y, et al. Feature-based online representation algorithm for streaming time series similarity search. Int J Patt Recogn Artif Intell, 2020, 34: 2050010
5 Hu Y, Ren P, Luo W, et al. Multi-resolution representation with recurrent neural networks application for streaming time series in IoT. Comput Netw, 2019, 152: 114–132
6 Keogh E, Lin J, Fu A W, et al. Finding unusual medical time-series subsequences: algorithms and applications. IEEE Trans Inform Technol Biomed, 2006, 10: 429–439
7 Keogh E, Chakrabarti K, Pazzani M, et al. Dimensionality reduction for fast similarity search in large time series databases. Knowledge Inf Syst, 2001, 3: 263–286
8 Ren H, Liao X, Li Z, et al. Anomaly detection using piecewise aggregate approximation in the amplitude domain. Appl Intell, 2018, 48: 1097–1110