Fast Shapelet Discovery with Trend Feature Symbolization

Shichao Zhang, Xiangwei Zheng(✉), and Cun Ji

School of Information Science and Engineering, Shandong Normal University, Jinan, China
jicun@sdnu.edu.cn

Abstract. Time series classification (TSC) is a hot topic in data mining field in the past decade. Among them, classifier based on shapelet has the advantage of interpretability, high accuracy and high speed. Shapelet is a discriminative subsequence of time series, which can maximally represent a class. Traditional fast shapelet algorithm uses SAX to represent time series. However, SAX usually loses the trend information of the series. In order to solve the problem, a trend-based fast shapelet discovery algorithm has been proposed. Firstly, the method of trend feature symbolization is used to represent time series. Then, a random mask is applied to select the candidate shapelets. Finally, the best shapelet is selected. The experimental results show that our algorithm is very competitive.

Keywords: Shapelet · Trend feature · Symbolization · Time series classification

1 Introduction

Time series classification (TSC) is one of the classical and hot issues in time series data mining. Time series come from a wide range of sources, including weather prediction, malware detection, voltage stability assessment, medical monitoring, and network anomaly detection [1]. In general, time series $T = \{t_1, t_2, \ldots t_m\}$ is a series of the values of the same statistical index arranged according to the time sequence of their occurrence [2]. The main goal of TSC is to divide an unlabeled time series into a known class.

In the existing time series classification algorithms, shapelet based algorithms are promising. Shapelet is a sub-sequence in time series which can represent a class maximally. These sub-sequences may appear anywhere in the time series and are generally shorter in length. Compared with other TSC algorithms, shapelet based classification method has the advantages of high classification accuracy, fast classification speed and strong interpretability [3]. In order to improve the speed of shapelet dicovery, a fast shapelet algorithm (FS) which uses Symbolic Aggregate Approximation (SAX) representation is proposed by Rakthanmanon and Keogh. However, SAX only uses the mean of sequence to represent time series, and it may cause the loss of trend information of time series. To solve this problem, a fast shapelet discovery algorithm based on Trend SAX (FS-TSAX) is proposed in this work. The main contributions of this paper are as follows:

Y. Sun et al. (Eds.): ChineseCSCW 2020, CCIS 1330, pp. 352–363, 2021.
https://doi.org/10.1007/978-981-16-2540-4_26

(1) A new shapelet discovery algorithm is proposed, combining FS and TSAX. It solves the shortcoming that SAX is easy to lose the trend information of time series and improves the accuracy of shapelet classification.

(2) Experiments are conducted on different data sets to evaluate the performance of the proposed algorithm. Experimental results show that the accuracy of our algorithm is at a leading level.

The remainder of this paper is structured as follows. Section 2 gives some related works on shapelet based algorithms and TSAX. Section 3 gives some definitions about FS-TSAX algorithm. Section 4 introduces our proposed FS-TSAX algorithm. Experimental results are presented in Sect. 5 and our conclusions are given in Sect. 6.

2 Related Works

2.1 Shapelet Based Algorithms

Since the concept of the shapelet was first proposed in 2009 [4], algorithms based on Shapelet have been proposed in large numbers.

However, shapelet-based algorithms are complex and take a long time to train [5]. For this, Rakthanmanon and Keogh proposed fast shapelet algorithm (FS) [6]. It uses SAX to reduce the dimension of the original data and uses the mean of sequence to represent time series. Then random masking the SAX string and construct Hash table statistics scores. Finally select the best shapelet according to the scores.

In addition, Wei et al. [7] combined existing acceleration techniques with sliding window boundaries and used the maximum correlation and minimum redundancy feature selection strategy to select appropriate shapelets. To dramatically speed up the discovery of shapelet and reduce the computational complexity, a random shapelet algorithm is proposed by Renard et al. [8]. In order to avoid using online clustering/pruning techniques to measure the accuracy of similar candidate predictors in Euclidean distance space, Grabocka et al. proposed a new method denoted as SD [9], which includes a supervised shapelet discover that filters out only similar candidates to improve classification accuracy. Ji et al. proposed a fast shapelet discovery algorithm based on important data point [12] and a fast shapelet selection algorithm [15]. The former accelerated the discovery of shapelet through important data points. The latter was based on shapelet transformation and LFDPs identification of the sampling time series, and then select the sub-sequences between two non-adjacent LFDPs as candidate sub-sequences of shapelet.

2.2 Trend-Based Symbolic Aggregate Approximation (TSAX) Representation

The symbolic representation of time series is an important step in data preprocessing, which may directly leads to the low accuracy of data mining. SAX is one of the most influential symbolic representation methods at present. SAX is a discrete method based on PAA, which can carry out dimensionality reduction processing simply and mine time series information efficiently. The main step of SAX is dividing the original time series into equal length sub-sequences, and then calculate the mean value of each subsequences and use the mean value to represent the subsequences, that is PAA. Then the breakpoint is found in the breakpoint table with the selected alphabet size, and the mean value of the PAA computed subsequences is mapped to the corresponding letter interval, finally the time series is discretized into strings [10].

However, SAX uses the letters after the mapping of PAA to represent each subsequence after segmentation, which may lose important features or patterns in the time series and lead to poor results in subsequent studies. As shown in Fig. 1, the result of SAX string is the same between two time series with completely different trend information.

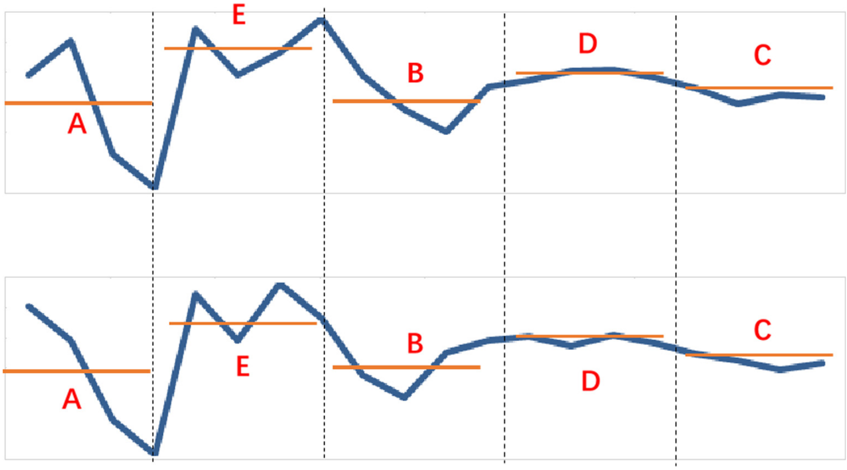

Fig. 1. Two time series with different trends get the same SAX string.

To solve the problem that SAX is easy to lose trend information of time series, Zhang et al. proposed a symbolic representation method based on series trend [11]. Specifically, after PAA is used on the time series, the sub-sequence with equal length are evenly divided into three segments, and the mean value of each segment is calculated respectively. Then a smaller threshold ε is defined and the size of the subsequence is calculated according to the formula (1). The letter “u” represents the upward trend, “d” represents the downward trend and “s” represents for horizontal trend. For example: if the mean of the first sub-sequence is less than the second one, represented as “u”, and the mean of the second sub-sequence is larger than the third one, represented as “d”, then this piecewise trend information is represented as “ud”, so the trend information of the time series is represented. TSAX is a combination of the SAX letters which represent trends.

$$Trend_{indicator} = \begin{cases} d, \overline{C_{i1}} - \overline{C_{i2}} > \varepsilon \\ s, \left|\overline{C_{i1}} - \overline{C_{i2}}\right| \le \varepsilon \\ u, \overline{C_{i1}} - \overline{C_{i2}} < \varepsilon \end{cases} \tag{1}$$

3 Definition

Definition 1: $T = \{t_1, t_2, \ldots t_m\}$ is a time series which contains an ordered list of numbers. Each value t_i can be any finite number and assume that m is the length of T.

Definition 2: S is a continuous sequence on time series, which can be expressed by formula (2). Where l is the length of S, i is the start position of S.

$$S = T_i^l = \{t_i, t_{i+1}, \ldots, t_{i+l-1}\} \tag{2}$$

Definition 3: Time series dataset D is a set of N time series, each of which is m in length and belongs to a specific class. The class number in D is C.

Definition 4: *(dist (T, R))* is a distance function, whose input is two time series $T = \{t_1, t_2, \ldots t_m\}$ and $R = \{r_1, r_2, \ldots r_m\}$. It returns a non-negative value. This paper uses Euclidean distance, and its calculation method is shown in formula (3).

$$dist(T, R) = \sqrt[2]{\sum_{i=1}^{m} (t_i - r_i)^2} \tag{3}$$

Definition 5: The distance between the subsequence S and the time series $T(subdist(T, S))$ is defined as the minimum distance between subsequence S and any subsequence of T of the same length as subsequence S. It is a distance function, which inputs time series T and sub-sequence S, returns a non-negative value. Intuitively, this distance is the distance between S and the best matching point at a certain position in T, as shown in Fig. 2, and its calculation method is given by formula (4).

$$SubDist(T, S) = \min(dist(T_1^l, S), dist(T_2^l, S), \ldots, dist(T_{m-l+1}^l, S)) \tag{4}$$

Definition 6: Entropy is used to indicate the level of clutter in a dataset. The entropy of dataset D shown in formula (5). Where D are datasets, C are different classes, and p_i is the proportion of time series in class i.

$$e(D) = -\sum_{i=1}^{c} (p_i \log p_i) \tag{5}$$

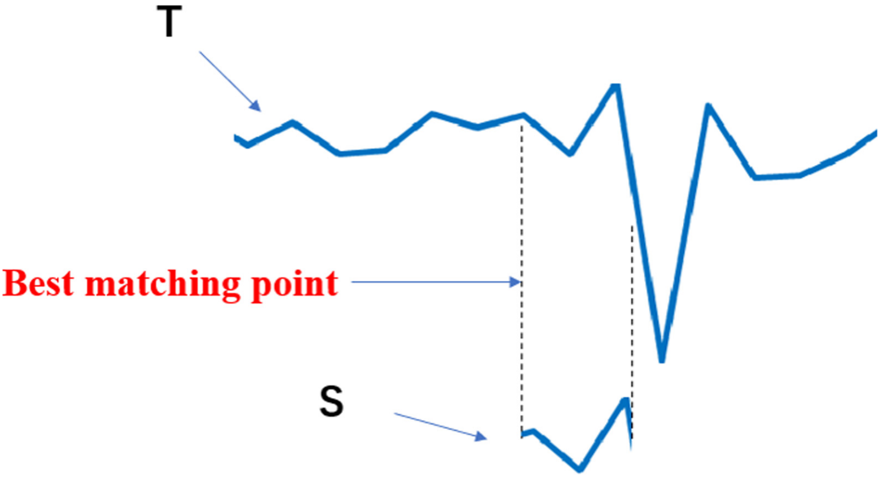

Fig. 2. Best matching point.

Definition 7: Information gain represents the degree of uncertainty reduction in a dataset under a partition condition. For a spilt strategy, the information gain calculation method is shown in Formula (6).

$$gain(TSAX) = e(D) - \frac{n_C}{n}e(D_C) - \frac{n_N}{n}e(D_N) \tag{6}$$

Definition 8: Optimal Split Point (*OSP*). When the information gain obtained by splitting at a threshold is larger than any other point, this threshold (distance value) is the *optimal split point*.

4 Fast Shapelet Discovery Algorithm Based on TSAX(FS-TSAX)

4.1 Overview of the Algorithm

The TS-TSAX algorithm is shown in Algorithm 1 and Fig. 3. Figure 3(a) shows the four processes of the FS-TSAX algorithm: (1) Generating TSAX words (Line 1–Line 3 in Algorithm 1); (2) Random masking of TSAX Words (Line 5–Line 7 in Algorithm 1); (3) Choose the top-k TSAX words with highest scores (Line 9 in Algorithm 1); (4) Find the best shapelet (Line 17–Line 21 in Algorithm 1).

Algorithm 1 *FastShapeletBasedOnTSAX*

Input: Time series dataset D
The number of iterations r
TSAX candidate number k

Output: *shapelet*

```
[TS,Label]←ReadData(D) // Read the time series and their labels
for (length = 1, length<m, length++)
    TSAXList ← creatTSAXList (TS,length) //creat TSAXList
    S ← {}
    for (i=1, i<r, i++)
        Calculate ← RandomMask (TSAXList,TS) // random masking TSAX words and calculating score
        S ← New S (S,Calculate)       //update scores
    end for
    TSAXCand ← FindTopKTSAX (SList,Calculate,k,r) //choose the best top-k shapelets
    Candshapelets ← Remap (TSAXCand,TS) // Select candidate shapelets

    Gain ← Inf, Gap ← 0
    for (i=1, i<Candshapelets, i++)
        candidate ← Candshapelets [i]
        Dist ← Euclidean distance (TS, candidate) // Calculate the distance between subsequence and time series
        [gain,gap] ← CalInfoGain(Dist) // Calculated in formation gain
        while (gain>Gain) ||
            ((gain==Gain)&&(gain>Gap))
            Gain ← gain
            Gap ← gap
            shapelet ← candidate //find shapelet
        end while
    end for
end for
```

Figure 3(b) shows the visual description of these four steps. Next, these four steps are described in detail.

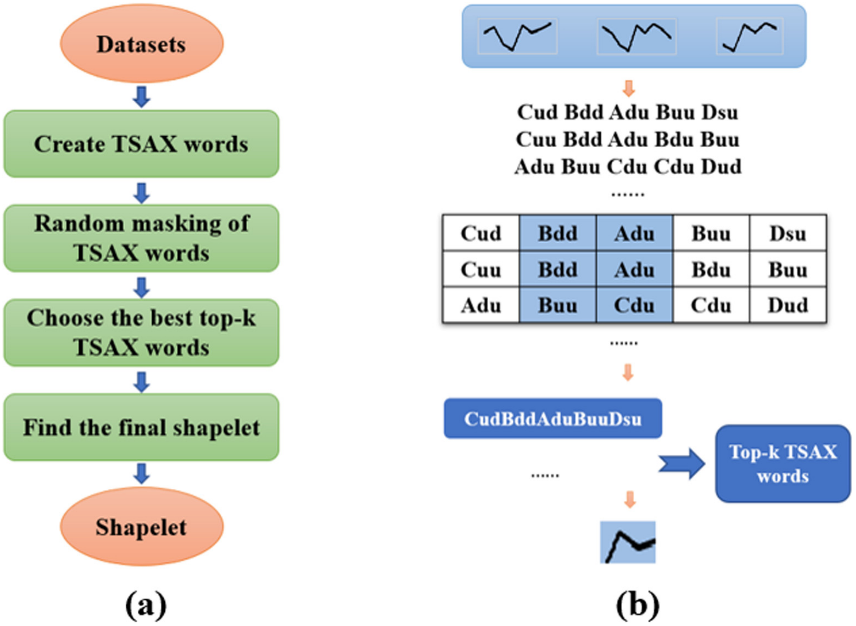

Fig. 3. The flow of FS-TSAX.

4.2 Create TSAX Words

For an original time series, after normalization, it is divided into equal length subsequences, then calculate the mean value of each segment and use the mean to find the corresponding letter in the breakpoint table, and a string is used to represent the time series, this process is SAX. After the time series is segmented with equal length by PAA, the sub-sequence segments are evenly divided into three segments. Calculate the mean of each segment and use formula (1) to get the trend letters. Then the trend letters of each segment with the SAX letter are combined, and the time series is represented in TSAX. Figure 4 shows an example of this process.

4.3 Random Masking TSAX Words

Two time series with similar real values may produce two different TSAX words just because of a minimal difference. Therefore, the best shapelet in the original time series may map to different TSAX words. The solution to this problem is to use random masking, which is the idea of projecting all the higher-dimensional TSAX words into smaller dimensions. The process of random masking TSAX words is as follows:

(1) Randomly select a character in a TSAX word.
(2) Select another character to generate the new TSAX word.

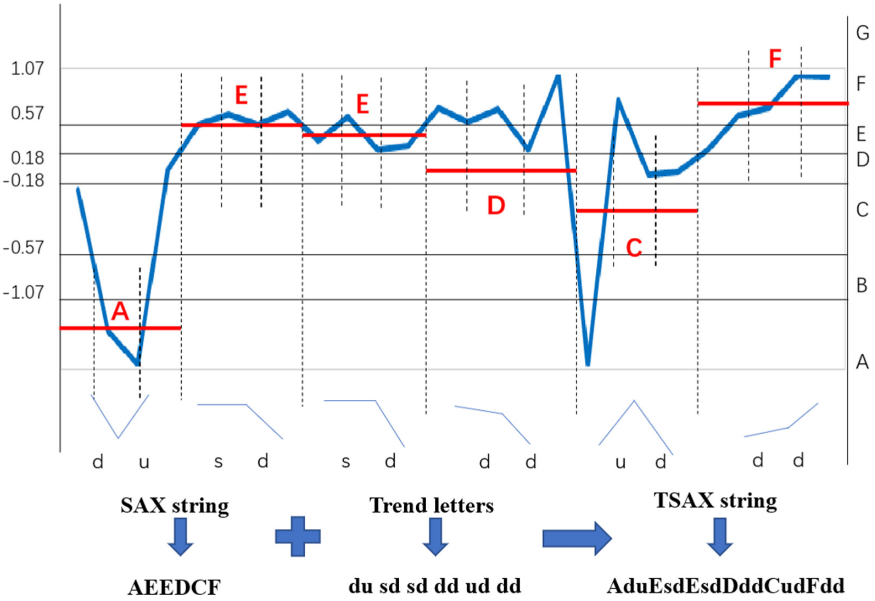

Fig. 4. Combine the SAX string "AEEDCF" and the trend letters "du sd sd dd ud dd" and get the TSAX string "AduEsdEsdDddCudFdd".

Typically, this process requires 10 iterations, which means masking 10 times TSAX words. After the TSAX words are randomly masked, the TSAX words and the words after masking are applied together for subsequent processing.

4.4 Choose the Best k TSAX Words

For a TSAX word, note 1 when it appears in a time series and 0 when it does not. By statistics of this information, training data set D can be divided into two sub-datasets: dataset D_C composed of time series containing the TSAX words, and dataset D_N composed of time series not containing the TSAX words. The information gain for this TSAX word can be calculated according to the following formula.

$$gain(TSAX) = e(D) - \frac{n_C}{n}e(D_C) - \frac{n_N}{n}e(D_N) \tag{7}$$

Where D is the dataset, D_C is the data set formed by the time series containing the TSAX words, D_N is the dataset formed by the time series without the TSAX words, n, n_C, and n_N are the number of time series contained in each of the three datasets. After calculating the information gain of TSAX, we find the best k TSAX words which has the best information gain and obtain the final shapelet.

4.5 Discover the Best Shapelet

Each TSAX word represents a corresponding time series, so after getting the top-k TSAX words, the corresponding relationship between TSAX words and time sequence can be used to get the corresponding sub-sequence. Then, we can find the final shapelet from the corresponding sub-sequence.

The finall shapelet is the one with the greatest information gain among the subsequences corresponding to the top k TSAX words. If there are multiple subsequences with the maximum information gain, the sub-sequence with the maximum clearance is selected [11].

5 Experiments and Evaluation

5.1 Datasets

UEA&UCR time series classification warehouse is an important open source data set in the field of time series data mining. In this chapter, we select 12 datasets from it for comparative experiments [13]. These data sets are set in "arff" format and each dataset sample carries a category label. Table 1 shows the information of these datasets.

Table 1. Dataset.

Data sets	Number of classes	Size of training dataset	Size of test dataset	Length of time series
Beef	5	30	30	470
ChlorineConcentration	3	467	3840	166
Coffee	2	28	28	286
ECG200	2	100	100	96
FaceFour	4	24	88	350
Haptice	5	155	308	1092
ItalyPowerDemand	5	67	1029	24
OSULeaf	6	200	242	427
Trace	2	810	3636	500
TwoLeadECG	2	23	1139	82
WordSynonyms	25	267	1139	82
Yoga	2	60	61	637

5.2 Effect of the Number of TSAX Segments

To verify the effect of the number of segments, we compared the classification accuracy on different data sets when the number of TSAX segments is 2 and 3. The experimental results are shown in Fig. 5. One thing to explain, theoretically, the number of segments is artificially selected. But our code uses binary to symbolize the time series, and the int in Java is 32 bits, and representing a letter needs two bits. A TSAX word is 15 characters, consisting of 5 SAX words and 10 trend letters, so it needs to be represented with 30 bits, and the maximum value of segment number is 3. Therefore, we compared the influence on classification accuracy when the number of segments is 2 and 3. As shown in Fig. 5, the accuracy of FS-TSAX (three-segments) on the 10 data sets was higher than that of FS-TSAX (two-segments), the accuracy of FS-TSAX (three-segments) on the 1 data sets was lower than that of FS-TSAX (two-segments), and they are equally accurate on 1 data set. In general, FS-TSAX (three-segments) is more competitive than FS-TSAX (two-segments).

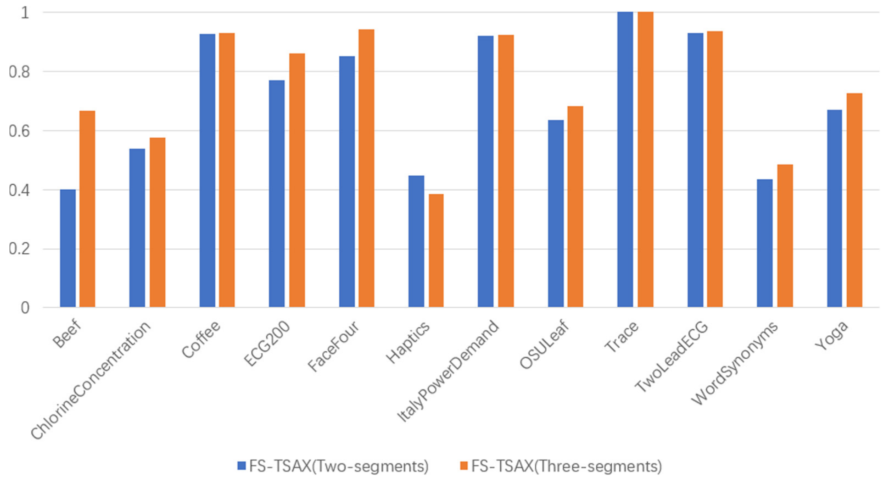

Fig. 5. Comparison of accuracy between FS-TSAX (three-segments) and FS-TSAX (two-segments).

5.3 Accuracy Comparison

In this section, we calculate the accuracy of our method and other three shapelet discovery algorithms. Table 2 shows the accuracy of these algorithms. The algorithm with the best result in each data set is shown in bold. The algorithms we selected are FS [6], SD [9] and the TSAX (Two-segments) algorithm mentioned in the previous section. As Table 2 shows, our approach gets the highest accuracy on 9 data sets, and the average rank is 1.25.

Table 2. Comparison of accuracy of different algorithms on different data sets.

	FS	FS-TSAX (Two-segments)	FS-TSAX (Three-segments)	SD
Beef	0.567	0.400	**0.667**	0.507
ChlorineConcentration	0.546	0.537	**0.577**	0.553
Coffee	0.929	0.927	0.929	**0.961**
ECG200	0.810	0.770	**0.860**	0.818
FaceFour	0.909	0.852	**0.943**	0.820
Haptice	0.376	**0.448**	0.386	0.356
ItalyPowerDemand	0.917	0.920	**0.922**	0.920
OSULeaf	0.678	0.636	**0.681**	0.566
Trace	**1.000**	**1.000**	**1.000**	0.965
TwoLeadECG	0.924	0.928	**0.936**	0.928
WordSynonyms	0.431	0.436	0.484	**0.625**
Yoga	0.695	0.671	**0.727**	0.625
Total wins	1	2	9	2
Average rank	2.5	2.75	1.25	2.58

6 Conclusion

The algorithms based on shapelet have attracted great attention in recent years. Shapelet discovery algorithm is the basis of shapelet transformation algorithm and shapelet learning algorithm, and it has various of acceleration strategies. We propose a shapelet discovery algorithm of FS-TSAX, which uses TSAX to represent time series, and can retain the trend information of time series well, and then carry out the process of shapelet discovery. Experiments on different data sets show that compared with other shapelet discovery algorithms, the accuracy of the proposed algorithm is at a leading level, especially in some time series with obvious trends.

References

1. Yan, W., Li, G.: Research on time series classification based on shapelet. Comput. Sci. **046**(001), 29–35 (2019)
2. Zhao, C., Wang, T., Liu, S., et al.: A fast time series shapelet discovery algorithm combining selective ex-traction and subclass clustering. J. Softw. **000**(003), 763–777 (2020)
3. Zhang, Z., Zhang, H., Wen, Y., Yuan, X.: Accelerating time series shapelets discovery with key points. In: Li, F., Shim, K., Zheng, K., Liu, G. (eds.) APWeb 2016. LNCS, vol. 9932, pp. 330–342. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45817-5_26
4. Ye, L., Keogh, E.: Time series shapelets: a novel technique that allows accurate, interpretable and fast classification. Data Mining Knowl. Discovery **22**(1–2), 149–182 (2011)

5. Bagnall, A., Lines, J., Bostrom, A., et al.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining Knowl. Discov. **31**(3) (2017)
6. Rakthanmanon, T., Keogh, E.: Fast shapelets: a scalable algorithm for discovering time series shapelets. In: Proceedings of the 2013 SIAM International Conference on Data Mining. Philadelphia: Society for Industrial and Applied Mathematics, pp. 668–676 (2013)
7. Wei,Y., Jiao, L., Wang, S., et al.: Time series classification with max-correlation and min-redundancy shapelets transformation. In: International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI), Beijing, China, pp. 7–12 (2015)
8. Renard, X., Rifqi, M., Detyniecki, M.: Random-shapelet: an algorithm for fast shapelet discovery. In: IEEE International Conference on Data Science and Advanced Analytics, pp. 1–10. IEEE (2015)
9. Grabocka, J., Wistuba, M., Schmidt-Thieme, L.: Fast classification of univariate and multivariate time series through shapelet discovery. Knowl. Inf. Syst. **49**(2), 1–26 (2015)
10. Lin, J., Keogh, E., Li, W., et al.: Experiencing SAX: a novel symbolic representation of time series. Data Mining Knowl. Discov. **15**(2), 107–144 (2007)
11. Zhang, K., Li, Y., Chai, Y., et al.: Trend-based symbolic aggregate approximation for time series representation. In: Chinese Control and Decision Conference (CCDC). Shenyang, pp. 2234–2240 (2018)
12. Ji, C., Zhao, C., Lei, P., et al.: A fast shapelet discovery algorithm based on important data points. Int. J. Web Serv. Res. **14**(2), 67–80 (2017)
13. Bagnall, A., Lines, J., Keogh, E., et al.: The UEA and UCR time series classification repository (2016). www.timeseriesclassification.com
14. Ji, C., Zhao, C., Liu, S., et al.: A fast shapelet selection algorithm for time series classification. Comput. Netw. **148**, 231–240 (2019)