# Hydrological Time Series Anomaly Mining based on Symbolization and Distance Measure

Dingsheng Wan, Yan Xiao, Pengcheng Zhang, Jun Feng, Yuelong Zhu, Qian Liu
College of Computer and Information, Hohai University, Nanjing, P.R.China 210098
Email: pchzhang@hhu.edu.cn; hhu_xiaoyan@163.com

***Abstract*—Large amount of hydrological data set is a kind of big data, which has much hidden and potentially useful knowledge. It is necessary to extract these knowledge from hydrological data set, which can provide more valuable hydrological information and be useful for future hydrological forecasting. Data mining based on time series is widely used currently. There are some techniques based on time series to extract anomaly. However, most of these techniques cannot suit big unstable data such as hydrological big data set. Some important problems are high fitting error after dimension reduction and low accuracy of mining results.**

**In this work we propose a new idea to solve the problem of hydrological anomaly mining based on time series. The idea combines time series symbolization with distance measure. It proposes Feature Points_Symbolic Aggregate Approximation (FP_SAX) to improve the selection of feature points, and then measures the distance of strings by Symbol Distance based Dynamic Time Warping (SD_DTW). Finally, the distance which we have got are sorted. A set of dedicated experiments are performed to validate our approach. The experimental data set is based on the water level data set obtained from Xiaomeikou gauge station in the Taihu Lake from 1956 to 2005. The results of experiments show that our approach has lower fitting error and higher accuracy.**

***Index Terms*—Hydrological Time Series; Data Mining; Pattern Representation; Distance Measure**

## I. Introduction

Hydrological data set are discrete records for hydrological process. They are huge, noisy, unstable and have poor correlation. Nowadays, hydrological modernization is frequently mentioned, which contains information collection, extraction, analysis and so on. During the hydrological process, it is indispensable to acquire valuable information and meaningful knowledge quickly from the numerous data. The rapid development of data mining provides a new approach for water resource management, hydrology and hydroinformatics research [19].

Data Mining [14] extracts potentially useful information which people is interested in and which is unknown in advance. Time series [3], which reflect the characteristic of the attribute value related with time, have large data scale, high dimension and update very quickly.

Hydrological time series data mining [19] is used to extract the unknown process that contains important information from hydrological data, which is valuable for hydrological forecasting and hydrological data analysis. Abnormal hydrological time series is the data object which is obviously inconsistent with the universal rule of the hydrological phenomenon. The research of abnormal hydrological time series data mining is still at the starting stage recently.

There are many approaches for abnormal time series data mining. Most of them have clear problems. For example, the approach based on immunology [4] cannot apply to diverse data. Computational efficiency of Support Vector Machine(SVM) [8] is high, and its theory and modeling process are very complex which can only be adopted by experts. The accuracy of TSA-tree [9] is low. To solve these problems, this paper puts forward a new approach which is based on Extended Symbolic Aggregate Approximation(ESAX) [16] and Dynamic Time Warping(DTW) [17].

In summary, the contributions of this paper are as follows:

- Approaches of selecting feature points (extreme points, minimum or maximum) are added into ESAX in pattern representation to reduce dimension, which is called FP_SAX. FP_SAX looks for new feature points which are more representative to replace the maximum and minimum proposed in ESAX. In FP_SAX, feature points consist of the following three parts: the beginning and ending points, the extreme feature points and the piecewise average feature points.
- This paper achieves a first combination of symbolic process and SD_DTW which is based on the distance between each symbol and DTW. A good mining result is acquired by this combination.
- A set of dedicated experiments have been conducted to validate our approach. Many suitable approaches are used in the experiment, such as Lagrange interpolation, data compression radio and so on. Experimental results show the low fitting error, acceptable time complexity and accuracy of our approach.

The rest of this paper is organized as follows. Section 2 reviews related work and discusses some background materials about time series data mining. In Section 3, the anomaly mining approach is proposed. Firstly, based on ESAX, we improve the approach of selecting feature points in symbolization, and then FP_SAX is proposed. Secondly, we measure the distance of strings according to the distance of DTW, which is called SD_SAX. The experimental validation of new approach is performed in Section 4. Finally, Section 5 offers some conclusions and suggestions for future work.

Pengcheng Zhang is Corresponding Author

978-1-4799-5057-7/14 $31.00 © 2014 IEEE
DOI 10.1109/BigData.Congress.2014.56

## II. Related work

Because of the large amounts and high dimensionality of time series data, using original data set pays abundant time and space cost. Dimensionality reduction techniques, also called sequence feature extraction, can translate big data to small data. After dimensionality reduction, we measure the distance of strings. In the following, we review related work about pattern representation and similarity measurement of time series.

### A. Pattern Representation

So far there have been four basic pattern representation approaches to extract sequence feature. These approaches are as follows: Piecewise Linear Representation (PLR) [7], Frequency Domain Representation (FDR) [2], Singular Value Decomposition (SVD) [10], Symbolic Aggregate Approximation (SAX) [13].

Keogh et al. [11] proposed Piecewise Aggregate Approximation (PAA). This approach divides original time series into several segments with equal length. The mean value of each segment is used as the feature of the segment. Then the original time series is represented by the feature of the segment. After that, Keogh proposed PLR [7]. In PLR, after being divided into several segments, time series is represented by end-to-end segments. As we all know, the number of segments has effect on the level of compression. But there is no criterion for the choice of the number. At the same time, PLR cannot applicable to nonlinear sequence.

There are two typical FDR approaches: Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). DFT appears in the field of digital signal processing early. Then it is proposed by Agrawal [1] again to apply to similarity searching. DFT allows a good dimension reduction, and measures the distance between each point in $k$-dimensional space by Euclidean distance. But some important extreme points are missed. In DWT [22], time series is analyzed by translation transformation and stretching transformation. The dimensionality is reduced without losing important points. However, both of them cannot apply to weighted Euclidean Distance.

SVD is a significant matrix distributing approach in linear algebra. SVD [10] transfers a group of given correlated variables into another group of uncorrelated variables whose variances are in a descending order, then generates a coordinate axis by mathematical manipulation. Due to the difference of the variance of each axis, original time series can be represented by several coordinate coefficient whose variance is in the top. Then the purpose of reducing dimensions is accomplished. However, original time series loses basic physical significance after SVD [6]. In addition, a reducing dimension progress involves global transformation of all data, which is relative complex.

SAX [13], which is the most representative symbolic approach, is proposed by Keogh. It is based on PAA for dimensionality reduction that minimizes dimensionality by the mean values of equal sized frames [16]. Then the results are turned into SAX symbols. However, SAX causes high possibilities to lose some points that may contain important information, such as the max/min points and extreme points. Finally, Lkhagva [16] put forward ESAX (Extended SAX) in 2006. ESAX overcomes the problem of missing important points exiting in SAX, which has a good performance in economic time series data mining. However, some of the feature points are missed in both of them.

### B. Similarity measurement

After reducing dimensionality, it is time to do similarity measurement. In many fields, distance measurement is similar with similarity measurement. Furthermore, distance between two objects is easy to compute. Therefore, similarity measurement is always replaced by distance measurement. There are two classical means: Euclidean Distance [21] and Dynamic Time Warping (DTW) [17].

Euclidean Distance [21], as the most widely used distance measurement, is simple, intuitive and easy to calculate. However, it is applicable to the time series with equal length only. And it is susceptible by noise and short time fluctuation.

Comparing with Euclidean Distance, DTW [17] is able to measure the distance without point-to-point correspondence when the timeline of time series has compand and bend. It also applies to the time series with different length.

## III. FP_SAX and SD_DTW

Hydrological data is nonlinear, huge and has poor correlation. The above proposed approaches cannot meet hydrological time series anomaly mining demand. In order to solve these problems, this paper proposes a new approach which contains FP_SAX and SD_DTW that have a good dimension reduction, low fitting error and relatively high accuracy.

### A. *ESAX*

ESAX is proposed to overcome the problem of losing critical and extreme points in SAX [13]. In ESAX, the sequence must conform to standard normal distribution. But hydrological data set is random and does not conform to standard normal distribution. So before symbolization, the hydrological sequence should be standardized. The specific steps of ESAX are as follows: [16]

**a)** Standardize original sequence $C$ to $C'$. $u$ and $v$ separately represent the mean value and standard deviation of this sequence:

$$c'_i = \frac{c' - u}{v} \tag{1}$$

**b)** Translate $C'$ whose length is $n$ to $X$ whose length is $k$ by PAA:

$$x_i = \frac{k}{n} \sum_{j=\frac{n}{k}(i-1)+1}^{\frac{n}{k}i} c'_j \tag{2}$$

**c)** Divide $X$ into $a$ equiprobable spaces. Based on the values of $x_1, x_2, ..., x_k$, the sequential values in the same probability space are represented by one symbol. The total number of

symbols is $a$. Then we get a symbol string whose length is $k$. Table I is the division of equal probability interval [12].

| $\beta$\a | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|
| $\beta_1$ | -0.43 | -0.67 | -0.84 | -0.97 | -1.07 | -1.15 | -1.22 | -1.28 |
| $\beta_2$ | 0.43 | 0 | -0.25 | -0.43 | -0.57 | -0.67 | -0.76 | -0.84 |
| $\beta_3$ | | 0.67 | 0.25 | 0 | -0.18 | -0.32 | -0.43 | -0.52 |
| $\beta_4$ | | | 0.84 | 0.43 | 0.18 | 0 | -0.14 | -0.25 |
| $\beta_5$ | | | | 0.97 | 0.57 | 0.32 | 0.14 | 0 |
| $\beta_6$ | | | | | 1.07 | 0.67 | 0.43 | 0.25 |
| $\beta_7$ | | | | | | 1.15 | 0.76 | 0.52 |
| $\beta_8$ | | | | | | | 1.22 | 0.84 |
| $\beta_9$ | | | | | | | | 1.28 |

TABLE I
THE DIVISION OF EQUAL PROBABILITY INTERVAL BASED ON THE NUMBER OF SYMBOLS (FROM 3 TO 10)

**d)** Find maximum and minimum of each PAA subsection $x_i$. Translate them separately into symbols $S_{max}$ and $S_{min}$. Remain their positions $P_{max}$ and $P_{min}$ at the same time. Then calculate the mean value of each subsection whose middle position is as follows(calculated from both the beginning position $S_k$, and the ending position $E_k$ on the time axis):

$$P_{mid} = \frac{S_k + E_k}{2} \tag{3}$$

**e)** The three symbols of each subsection $x_i$ can be represented by the following equation:

$$< S_1, S_2, S_3 >= \begin{cases} < S_{max}, S_{mid}, S_{min} > if P_{max} < P_{mid} < P_{min} \\ < S_{min}, S_{mid}, S_{max} > if P_{min} < P_{mid} < P_{max} \\ < S_{min}, S_{max}, S_{mid} > if P_{min} < P_{max} < P_{mid} \\ < S_{max}, S_{min}, S_{mid} > if P_{max} < P_{min} < P_{mid} \\ < S_{mid}, S_{max}, S_{min} > if P_{mid} < P_{max} < P_{min} \\ < S_{mid}, S_{min}, S_{max} > otherwise \end{cases} \tag{4}$$

In ESAX, time series are divided equally. All changes are considered in the same way whether they are slow or not. In fact, these two situations represent different significance in time series, especially in hydrological phenomena. If merely using maxima and minima of every segment as feature points, we are unable to make sure that every point which has effect on the series is included.

### B. *FP_SAX*

*1) Dimension reduction approaches based on extreme points*

In the task of hydrological time series data mining, the peaks and valleys in a time series are often particularly important to analyze the transformation of hydrological phenomena for a certain period time. So peaks and valleys are more special than other sequence points, and these extreme points must be retained as much as possible. Therefore, to keep extreme points such as peaks and valleys, dimension reduction approaches based on extreme points are indispensable.

**Definition 1: Extreme points.** For time series $X = (x_1, x_2, ..., x_m)$, if one of the following conditions is met, we call $x_i$ the extreme point:

$$(1) x_i \geq x_{i-1} and x_i \geq x_{i+1};$$

$$(2) x_i \leq x_{i-1} and x_i \leq x_{i+1}.$$

This approach can retain all extreme points of the series completely. But its ability to reduce dimension depends on changes of the series, because it just simply considers the relative size between the previous sequence and the later one, and has no macroscopic observation of the entire sequence.

Another widely used dimension reduction approach – PAA [20], which is simple and easy to implement, has a good effect on the sequence which changes gently. Its core idea is to fix the width of window to keep the consistency. Then it calculates piecewise average of each window. The width of each window is demanded to match. If the width of the window is too wide, it can't reflect changes of sequences in time. If too narrow, the data compression is small relatively. It cannot reach the purpose of dimension reduction. Therefore, improper selection of the window will have a big impact on subsequent mining.

In order to solve the problems existed in extreme points and PAA dimension reduction approaches which are used for feature extraction of time series, we improve the traditional technology of extreme points and ESAX Symbolic approaches, then propose FP_SAX which can meet the needs of dimension reduction in the special type of hydrological time series.

*2) The selection of feature points*

FP_SAX is still based on SAX. Besides the maximum and minimum proposed in the ESAX, we look for new feature points which are more representative to replace the maximum and minimum, to make sure that the important information of the sequence won't be lost. The symbolic process according to these new feature points is able to preserve important feature of original sequence.

In FP_SAX, feature points consist of the following three parts: the beginning and ending points, the extreme feature points which keep extreme time period and the piecewise average feature points containing certain number of extreme points.

**Definition 2: Extreme feature points.** [23] For time series $X = (x_1, x_2, ..., x_m)$, if meeting the following two conditions, we call $x_i$ extreme feature points:
(1)$x_i$ must be extreme value point of the sequence;
(2)$x_i$ keeps extreme time period (the distance between the previous extreme point and the later one). The ratio, which is obtained through dividing extreme time period by the length of this sequence, must be not less than the threshold value B. According to the length of original time series and domain knowledge, the value of $B$ is usually between 0.001 and 0.1.

**Definition 3: Average feature points.** Time series $X = (x_1, x_2, ..., x_m)$ has $k$ extreme feature points. Divide the sequence into subsequences to make sure that each subsequence contains $N$ extreme feature points. The mean value of each subsequence is called average feature point.

The minimum value of $N$ is 1, which means that this subsequence only contains one extreme point. The maximum value of $N$ is $k$ which is the number of all extreme feature points of this sequence. At this time, the sequence has only one average feature point that is the mean value of all sequence points.

Average feature points break the model that PAA dimension reduction approaches must be in the fixed window size. The width of window is decided by the change of sequence. If a subsequence contains many feature points, frequent fluctuations of this subsequence are illustrated and this subsequence has an effect on the whole sequence. So it must be recorded in detail. Then we should narrow the width of the window. If a sequence contains less feature points, which shows that this sequence changes gently, we should enlarge the width of window, which plays a significant role in the improvement of data compression ratio. Average feature points can both achieve the purpose of dimension reduction, and remain important characteristics of the sequence.

After selecting suitable feature points, according to abscissa ascending order size of feature points, the total number of symbols is supposed to be determined. According to table I [12] which provides the dividing principle of normal distribution equal probability interval, each feature point is mapped to the matched symbol interval. Then symbols are obtained, and original sequence is translated into one string.

The algorithm of FP_SAX is specified in Algorithm 1. Lines 1 to 7 standardize original sequences. Lines 8 to 12 keep extreme points. In lines 14-17, it determines the extreme feature points and saves them. Finally, feature points are transformed to symbols according to Table I and equation (4).

### C. SD_DTW

After FP_SAX, it is time to do similarity measurement. DTW is more accurate than Euclidean distance and applies to the compand of the time shaft, so some distance measurement is developed from it. Based on the idea of DTW, combining the distance between the FP_SAX symbols, SD_DTW is proposed to solve the problem of distance measurement.

SD_DTW approach describes the distance between each symbol through a matrix, of which $i$ and $j$ respectively represent rows and columns. The element of matrix is as follows: [5]

$$dis[i][j] = \begin{cases} o, if \mid i - j \mid \leq 1 \\ \beta_{max(i,j)-1} - \beta_{min(i,j)}, otherwise \end{cases} \quad (5)$$

The value of $\beta_n$ is in the reference Table I [12].

For example, when total number of symbols a is 5, A, B, C, D, E are used to represent the original time series. Then the distance between each symbol is shown in Table II.

**Algorithm 1** Dimensionality Deduction of Time Series based on FP_SAX

**Require:** Original time series X=$(x_1, x_2, ..., x_i, ..., x_n)$, threshold value $w$, number of extreme feature points $k$;

**Ensure:** The string $S = (s_1, s_2, ..., s_i, ..., s_m)$;

1: **for** $i = 0$ *to* $n$ **do**
2: $x_i \rightarrow a[i]$;
3: *Calculate the mean value and standard deviation u*;
4: **end for**
5: **for** $i = 0$ *to* $n$ **do**
6: $(a[i] - mean)/u \rightarrow a[i]$;
7: **end for**
8: **for** $i = 0$ *to* $n$ **do**
9: **if** ((a[i]>a[i+1]) and (a[i]>a[i-1])) or ((a[i]<a[i+1]) and (a[i]<a[i-1])) **then**
10: $a[i] \rightarrow b[j]; j++$;
11: **end if**
12: **end for**
13: **for** $i = 1$ *to* $j - 1$ **do**
14: **if** (b[i+1]-b[i-1]>w*n) **then**
15: $b[i] \rightarrow c[p].value$;
16: $i \rightarrow c[p].location; p++$;
17: **end if**
18: **if** (p == k) **then**
19: *Calculate the mean value and coordinate*;
20: *continue*;
21: **end if**
22: **end for**
23: **for** $i = 0$ *to* $m$ **do**
24: $s[i]$ = *corresponding symbols of mean and extreme feature points*;
25: $cout \ll s[i]$;
26: **end for**

| | A | B | C | D | E |
|---|---|---|---|---|---|
| A | 0 | 0 | 0.59 | 1.09 | 1.68 |
| B | 0 | 0 | 0 | 0.5 | 1.09 |
| C | 0.59 | 0 | 0 | 0 | 0.59 |
| D | 1.09 | 0.5 | 0 | 0 | 0 |
| E | 1.68 | 1.09 | 0.59 | 0 | 0 |

TABLE II
THE DISTANCE BETWEEN EACH SYMBOL WHEN A=5

**Definition 4: Symbol Distance based Dynamic Time Warping (SD_DTW).** Two strings $S = (s_1, s_2, ..., s_{n1})$ and $T = (t_1, t_2, ..., t_{n2})$, whose length are separately $n_1$ and $n_2$, are arranged by time. An $m*n$ matrix $A$ is constructed to represent DTW distance of the two strings. The element $a_{ij}$ in matrix A is the distance between $s_i$ and $t_j$, $d(s_i, t_j)$, the equation is shown as (5).

There is a set $W = w_1, ..., w_k, ..., w_K (w_k = d(x_i, y_j))$ containing a group continuing matrix elements. $W$ is the bent lane of $S$ and $T$, which complies with the following principles: [18]

(i) *Boundary:* both beginning point and ending point of the bent lane are located in back-diagonal of the relational matrix and separately the beginning/ending point of two time series are $w_1 = d(x_1, y_1), w_k = d(x_m, y_n)$.

(ii) *Continuity:* any two points of the bent lane must be adjacent elements or diagonal adjacent elements of relational matrix. If $w_k = d(x_a, y_b), w_{k-1} = d(x_{a'}, y_{b'})$, then $a - a' \leq 1, b - b' \leq 1$.

(iii) *Monotonicity:* all points in the bent lane should follow monotonicity. If $w_k = d(x_a, y_b), w_{k-1} = d(x_{a'}, y_{b'})$, then $a - a' \geq 0, b - b' \geq 0$.

Following is the equation of DTW distance: [18]

$$D(1,1) = d(x_1, y_1);$$
$$D(i,j) = d(x_i, y_j) + min\{D(i-1, j), D(i, j-1), D(i-1, j-1)\} \quad (6)$$

*D(i,j)* is cumulative distance, which is the sum of minimum value of this point and minimum bent path in the upper left. This is a kind of dynamic planning approach based on cumulative distance matrix to calculate DTW distance.

The algorithm of SD_DTW is shown in Algorithm 2. In lines 2-4 and 5-7, the algorithm separately gives the distance of first line and column. Lines 8-11 calculate cumulative distances.

**Algorithm 2** SD_DTW

**Require:** String $X$ and String $Y$, whose length are separately $m$ and $n$;
**Ensure:** Distance of String $X$ and String $Y$ based on DTW;

```
1:  v[0][0] = dis(x[0], y[0]);
2:  for i = 0 to m do
3:      v[i][0] = v[i − 1][0] + dis(x[i], y[0]);
4:  end for
5:  for j = 0 to n do
6:      v[0][j] = v[0][j − 1] + dis(x[0], y[j]);
7:  end for
8:  for j = 0 to n do
9:      for i = 0 to m do
10:         v[i][j] = dis(x[i], y[j]) + min{v[i−1][j], v[i][j−1], v[i−1][j − 1]};
11:     end for
12: end for
13: return v[m − 1][n − 1]
```

The time complexity of DTW is $O(n_1 * n_2)$, $n_1$ and $n_2$ are separately lengths of strings.

For instance, we want to calculate DTW distance between S=ABDC and T=DCADBE. The total number of symbols a is 5. Then DTW distance between $S$ and $T$ is 1.68, and the red figure is the best winding path.

## IV. Experimental evaluation

In this section, we conduct a set of experiments to show the accuracy and usability of our new approach by comparing it with ESAX [15] which is proposed recently. The experiments are designed to answer the following three research questions:

| T \ S | A | B | D | C |
|---|---|---|---|---|
| D | 1.09 | 1.59 | 1.59 | 1.59 |
| C | 1.68 | 1.09 | 1.09 | 1.09 |
| A | 1.68 | 1.09 | 2.18 | 1.68 |
| D | 2.77 | 1.59 | 1.09 | 1.09 |
| B | 2.77 | 1.59 | 1.59 | 1.09 |
| E | 4.45 | 2.58 | 1.59 | 1.68 |

TABLE III
The DTW distance between X and Y

- REQ 1: What is the difference of fitting error between the two approaches under different data compression ratios?
- REQ 2: How much execution time is required?
- REQ 3: How about the accuracy of our approach?

**Experimental setup:** To solve REQ 1, fitting error is gained by interpolation approaches based on daily water level data from Xiaomeikou gauge station in 2006. The data compression radio is changed. We compare fitting errors under different data compression radios. For REQ 2, we separately select water level data in the same period of 2 months (July and August), 3 months (June to August), 4 months (June to September), 5 months (June to October), 6 months (May to October) and 7 months (April to October) from 1956 to 2005. Then the average execution time of each subsequence is calculated by two approaches respectively. In REQ 3, water level data in May and June from 1956 to 2005 are used to get the first six largest distances by the two approaches, which are the abnormal patterns.

Software and hardware environment are presented in Table IV.

| Num | Property | Parameters |
|---|---|---|
| 1 | Processor | Intel(R)Core 2 |
| 2 | CPU | 2.40GHz |
| 3 | Memory | 2G |
| 4 | OS | Windows Vista |
| 5 | Software | VC++6.0 |

TABLE IV
Software and hardware environment

**Experimental process and results:**

*REQ 1: What is the difference of fitting error between the two methods under different data compression ratios?*

A dimensionality reduction approach can be judged by fitting error between original sequence and the one after reducing dimension. Under the same data compression radio, the smaller fitting error means that this dimensionality reduction approach is better.

To solve this question, two expressions will be proposed, and their definitions are described as follows:

*Data compression radio*: Translate time series $X = (x_1, x_2, ..., x_N)$ to a new series $X' = (x'_1, x'_2, ..., x'_n)(n<N)$ by reducing dimension, which means that the dimensionality is changed from $N$ to $n$. Then data compression radio is shown as the following equation:

$$\frac{N-n}{N} * 100\% \tag{7}$$

*Fitting error*: Translate time series $X = (x_1, x_2, ..., x_N)$ to a new series $X' = (x'_1, x'_2, ..., x'_n)(n<N)$ by reducing dimension. Then restore $X'$ into another sequence $Y' = (y'_1, y'_2, ..., y'_N)$ which has the same dimension with $X$ by the approach of interpolation. Thus the fitting error between $X'$ and $X$ is defined as follows:

$$\delta = \sum_{i=1}^{N} |x_i - y_i| \tag{8}$$

Interpolation approaches are used to gain the fitting error. There are three basic interpolation approaches: linear interpolation, least square interpolation approach and Lagrange interpolation. Linear interpolation is applied to linear data or polynomial function. Hydrological data has big fluctuation and is nonlinearity as a whole. So the approach of linear interpolation is not a good choice. The function fitted by least square interpolation is not necessarily passing sample points. But hydrological data needs to remain the information of sample points when fitting, which means that this approach is not applied to hydrological data. Lagrange interpolation requires new function to pass sample points, and has a better effect on nonlinear function than other approaches. Therefore, this part uses Lagrange interpolation to gain fitting error which is used to evaluate the property of dimensionality reduction approach, whose theory is as follows:

$T$ between $T_i$ and $T_{i+1}$ can be calculated by the first three points $T_{i-1}$, $T_i$ ,$T_{i+1}$, and also by the latter three points $T_i$, $T_{i+1}$, $T_{i+2}$. The interpolation formula of front three points is shown as equation 9, and the one of latter three points is as equation 10:

$$T = \frac{(t-t_i)(t-t_{i+1})}{(t_{i-1}-t_i)(t_{i-1}-t_{i+1})}T_{i-1} + \frac{(t-t_{i-1})(t-t_{i+1})}{(t_i-t_{i-1})(t_i-t_{i+1})}T_i + \frac{(t-t_i)(t-t_{i-1})}{(t_{i+1}-t_i)(t_{i+1}-t_{i-1})}T_{i+1} \tag{9}$$

$$T = \frac{(t-t_{i+1})(t-t_{i+2})}{(t_i-t_{i+1})(t_i-t_{i+2})}T_i + \frac{(t-t_i)(t-t_{i+2})}{(t_{i+1}-t_i)(t_{i+1}-t_{i+2})}T_{i+1} + \frac{(t-t_i)(t-t_{i+1})}{(t_{i+2}-t_i)(t_{i+2}-t_{i+1})}T_{i+2} \tag{10}$$

To improve the reliability of results, the mean value of interpolating points respectively from the front and latter three points is used for the final interpolating point.

We want to compare fitting error of the two approaches (ESAX in [15] and the approach in this paper) by Lagrange interpolation under the same data compression radio. The following are the specific experimental steps.

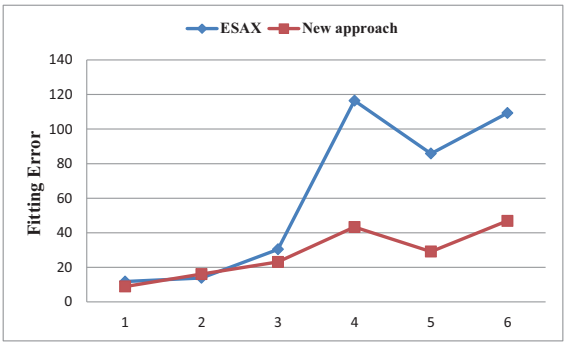


Fig. 1. The comparison of fitting error of ESAX and new approach

*Step 1*: Select the daily water level data of all the year round in 2006.

*Step 2*: Change the parameter in the algorithm of ESAX in [15]: the number of segments in PAA. Then we can get feature points under different data compression radios. Calculate the fitting error between the new sequence based on Lagrange interpolation and the original sequence.

*Step 3*: Change parameters in the algorithm of new approach: The threshold value $B$ which keeps extreme time period and the number of extreme feature points in a subsequence $N$. Then we can get feature points under different data compression radios. Calculate the fitting error between the new sequence based on Lagrange interpolation and the original sequence.

*Step 4*: List and draw the statistical graphics.

| Num | ESAX | | | New approach | | |
|---|---|---|---|---|---|---|
| | Dim | DCR | FE | Dim | DCR | FE |
| 1 | 274 | 24.9% | 11.767 | 270 | 26.0% | 8.908 |
| 2 | 157 | 57.0% | 13.882 | 158 | 56.7% | 16.096 |
| 3 | 85 | 76.7% | 30.399 | 84 | 76.9% | 23.107 |
| 4 | 42 | 88.5% | 116.457 | 42 | 88.5% | 43.284 |
| 5 | 36 | 90.1% | 85.822 | 38 | 89.6% | 29.079 |
| 6 | 30 | 91.8% | 109.275 | 29 | 92.0% | 46.876 |

TABLE V

THE FITTING ERROR OF ESAX AND NEW APPROACH (DIM REPRESENTS DIMENSION AFTER REDUCING,DCR REPRESENTS DATA COMPRESSION RADIO,FE REPRESENTS FITTING ERROR)

Table V and Figure 1 show that the two fitting errors are almost same when data compression radio is relatively small. With the increasing of compression radios, the fitting error also increases. However, the fitting error of our new approach is always lower than ESAX, especially when data compression radio is relatively large. Furthermore, the growth rate of fitting error of ESAX is higher than our approach. Therefore, when extracting the feature of time series, our approach is a better choice to obtain important feature and extract key feature points of the original sequence.

*REQ 2: How much execution time is required?*

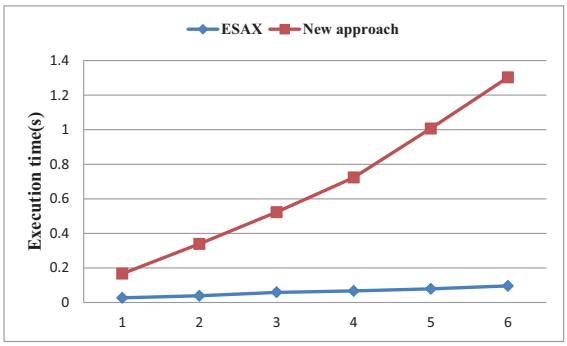


Fig. 2. The comparison of runtime between ESAX and new approach

The execution time is a key factor to estimate an algorithm. In this section, we compare the execution time of ESAX and the new approach along with the increase of the length of sequences. The experimental data is the daily water level data of Xiaomeikou gauge station from 1956 to 2005 in Taihu Lake.

| Num | Month | The length of subsequence | Total search length | ESAX | new approach |
|---|---|---|---|---|---|
| 1 | Jul. and Aug. | 62 | 3100 | 0.027s | 0.167s |
| 2 | Jun. to Aug. | 92 | 4600 | 0.039s | 0.339s |
| 3 | Jun. to Sep. | 122 | 6100 | 0.059s | 0.523s |
| 4 | Jun. to Oct. | 153 | 7650 | 0.067s | 0.723s |
| 5 | May to Oct. | 184 | 9200 | 0.079s | 1.007s |
| 6 | Apr. to Oct. | 214 | 10700 | 0.096s | 1.303s |

TABLE VI
THE EXECUTION TIME OF THE TWO APPROACHES

From Table VI and Figure 2, we can see that the execution time of the new approach is obviously higher than ESAX. The main reason is that ESAX adopts Euclidean distance which needs only one **for** loop to get the distance between strings. However, in our approach, the length of strings after symbolization is different. Consequently we adopt the approach based on DTW to measure precisely. The time complexity of our approach is $O(m * n)$($m$ and $n$ separately represent the length of two strings $X$ and $Y$) which spends most of time to measure distance between strings and is relatively high. In spite of higher time complexity of our approach, the execution time can still remain about one second when total length of the sequence is more than ten thousands, which is tolerated in most cases.

*REQ 3: How about the feasibility and accuracy of our approach?*

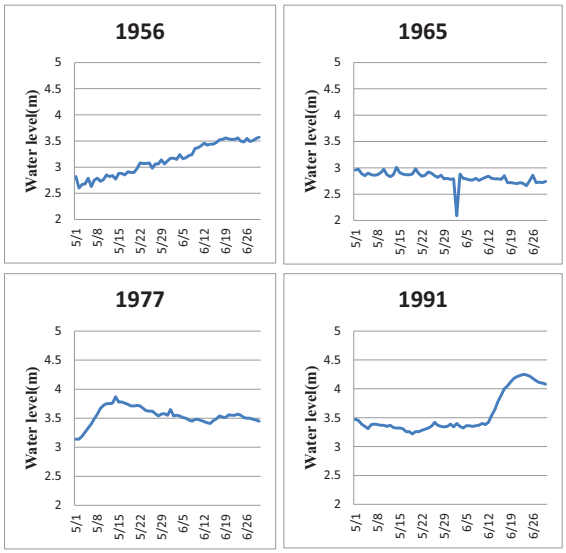


Fig. 3. The water level in May and June

We use daily water lever data of Xiaomeikou gauge station in Taihu Lake. The data are obtained from 1956 to 2005. We select the parameter values as follows. The sequence length in our experiment is 61. Consequently the threshold value $B$ is designated as 0.05. The number $N$, which is the number of extreme feature points in subsequence, is usually 4 or 5 that has little effect on the experimental results. The total number of symbols $a$ is 5. The top six largest distances (Top_5) are chosen as the results of abnormal mining.

| Results | ESAX | | New approach | |
|---|---|---|---|---|
| | Year | Distance | Year | Distance |
| 1 | 1977 | 602.91 | 1977 | 513.26 |
| 2 | 1991 | 588.20 | 1991 | 473.11 |
| 3 | 2002 | 560.50 | 1979 | 462.45 |
| 4 | 1960 | 545.75 | 1997 | 458.18 |
| 5 | 1972 | 539.69 | 1965 | 450.29 |
| 6 | 1959 | 536.20 | 1973 | 443.77 |

TABLE VII
THE MINING RESULTS OF THE SUBSEQUENCE IN MAY AND JUNE

Sequence abnormal pattern mining results of May and June are shown in Table VII. Firstly, according to hydrological characteristics and rules of season change in Taihu Lake, the water level in May is relatively stable and rises in June, such as in 1956 shown in Figure 3. However, there is an abnormal pattern in which the water level firstly rose then fell of May and June in 1977. Secondly, the water level of June in 1991 rose rapidly. Furthermore, the increasing range and speed is higher than any other year, which is also an abnormal pattern. The above two abnormal patterns can be extracted by both approaches. Finally, there is an obvious outlier which

is extracted by the approach proposed by us but ignored by ESAX. As we can see in figure 3, the water level of May and June has a glaring trough in 1965: the height of water is 2.04 meters on the 2nd of June that has a big fall compared with adjacent days, which is a conspicuous outlier. Besides the above three abnormal pattern results, other years shown in Table VII are not intuitively obvious abnormal patterns which need more further analyses by hydrologists.

From the experiments results describe above, our proposed approach is more accurate than ESAX and can mine a variety of different types of abnormal data. The feasibility and accuracy of our approach is validated. It is of great significance for the observation and study of hydrological phenomenon.

**Threats to validity:** Although experimental results reveal the accuracy and usability of our approach, there are still some threats to validity.

Firstly, the choice of *N*, which is the number of extreme feature points of each subsequence, and the threshold value B have a much effect on the accuracy of our approach. They are based on abundant experiments. Consequently, wrong choice may lead to the failure of experiments.

Secondly, our approach can only mine anomaly currently. In other words, it is unable to distinguish the anomaly (which is flood or which is drought).

## V. Conclusions and suggestions for future work

This paper studies the problem of the abnormal hydrology time sequence mining. Combining with the field of hydrology, we explore the effective and accurate approach of abnormal pattern mining. At present, the abnormal time pattern mining, especially in the field of hydrology, is based on distance measurement. These approaches need to calculate the distance between each pattern, whose time complexity is very high. In this work, we combine symbolization (FP_SAX) of time series with distance measurement (SD_DTW). Then a new approach that is suitable for feature extraction of hydrological time series and can dig out abnormal model quickly is found, which makes the mining results accurate and efficient.

There are some parts that remain to be improved in the future. Firstly, in FP_SAX algorithm, the minimum value of *N*, which is the number of extreme feature points of each subsequence, is 1, and the maximum value is the number of total extreme feature points. This paper estimates the value of *N* based on the experience of previous experiments. In the future we should consider a more scientific way of evaluation, which achieves the optimal value of *N*. Secondly, how to make the classification result of subsection with different length to be more suitable with feature extraction is also an interesting question. Finally, our approach has higher time complexity compare to ESAX. Although it is tolerable for most cases, new techniques are needed to reduce the time complexity.

## VI. Acknowledgements

The work is supported by the National Natural Science Foundation of China under Grant (Nos. 61370091, 51079040 and 61202097), Jiangsu Province Science and Technology Support Program Project under Grant (No.BE2012179), China Postdoctoral Foundation (Nos. 2012T50489 and 2011M500897), and Doctoral Fund of Ministry of Education of China (Grant No.20120094120009).

## References

[1] R. Agrawal, C. Faloutsos, and A. N. Swami, "Efficient similarity search in sequence databases," in *FODO*. Springer, 1993, pp. 69–84.

[2] P. Bigioi, M. Ciuc, S. Ciurel, P. Corcoran, Y. Prilutsky, E. Steinberg, and C. Vertran, "Classification system for consumer digital images using automatic workflow and face detection and recognition," Jun. 12 2012, uS Patent 8,199,979.

[3] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time series analysis: forecasting and control*. Wiley. com, 2013.

[4] P. Chaovalit, A. Gangopadhyay, G. Karabatis, and Z. Chen, "Discrete wavelet transform-based time series analysis and mining," *ACM Computing Surveys (CSUR)*, vol. 43, no. 2, pp. 6–37, 2011.

[5] A. J. Enright, S. Van Dongen, and C. A. Ouzounis, "An efficient algorithm for large-scale detection of protein families," *Nucleic acids research*, vol. 30, no. 7, pp. 1575–1584, 2002.

[6] P. Esling and C. Agon, "Time-series data mining," *ACM Computing Surveys (CSUR)*, vol. 45, no. 1, pp. 12–34, 2012.

[7] T.-c. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011.

[8] V. Gómez-Verdejo, J. Arenas-García, M. Lazaro-Gredilla, and A. Navia-Vazquez, "Adaptive one-class support vector machine," *Signal Processing, IEEE Transactions on*, vol. 59, no. 6, pp. 2975–2981, 2011.

[9] M. Gupta, J. Gao, C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, vol. 25, no. 1, pp. 1041–1060, 2013.

[10] E. Henry, J. Hofrichter *et al.*, "Singular value decomposition: application to analysis of experimental data," *Essential Numerical Computer Methods*, vol. 210, no. 6, pp. 81–138, 2010.

[11] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality reduction for fast similarity search in large time series databases," *Knowledge and information Systems*, vol. 3, no. 3, pp. 263–286, 2001.

[12] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*. ACM, 2003, pp. 2–11.

[13] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing sax: a novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107–144, 2007.

[14] K.-P. Lin and M.-S. Chen, "On the design and analysis of the privacy-preserving svm classifier," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 23, no. 11, pp. 1704–1717, 2011.

[15] Q. Liu, Y. Zhu, and P. Zhang, "Extended symbolic aggregate approximation based anomaly mining of hydrological time series," *Journal of Computer Application Research*, vol. 29, no. 12, pp. 4479–4502, 2012.

[16] B. Lkhagva, Y. Suzuki, and K. Kawagoe, "New time series data representation esax for financial applications," in *ICDE Workshops*, 2006, pp. 115–136.

[17] M. Müller, "Dynamic time warping," *Information Retrieval for Music and Motion*, vol. 4, pp. 69–84, 2007.

[18] R.-L. OUYANG, L.-L. Ren, and C.-H. Zhou, "Similarity search in hydrological time series," *Hohai University( Natural Sciences)*, vol. 38, no. 3, pp. 241–245, 2010.

[19] R. Ouyang, L. Ren, W. Cheng, and C. Zhou, "Similarity search and pattern discovery in hydrological time series data mining," *Hydrological Processes*, vol. 24, no. 9, pp. 1198–1210, 2010.

[20] P. Papapetrou, V. Athitsos, M. Potamias, G. Kollios, and D. Gunopulos, "Embedding-based subsequence matching in time-series databases," *ACM Transactions on Database Systems (TODS)*, vol. 36, no. 3, pp. 17–39, 2011.

[21] J. C. Russ, *The image processing handbook*. CRC press, 2010.

[22] J.-L. Starck, F. Murtagh, and J. M. Fadili, *Sparse image and signal processing: wavelets, curvelets, morphological diversity*. Cambridge University Press, 2010.

[23] H. Xiao and Y. Hu, "Data mining based on segmented time warping distance in time series database." *Journal of Computer Research and Development*, vol. 42, no. 1, pp. 72–78, 2005.