Sequential Pattern Mining and Matching Method with Its Application on Earthquakes

Jiang Zhu[1(✉)], Dechang Pi[1], Pan Xiong[2], and Xuhui Shen[2]

[1] College of Computer Science and Technology,
Nanjing University of Aeronautics and Astronautics,
29 Yudao Street, Nanjing, 210016 Jiangsu, People's Republic of China
zhujiang_09l0@yeah.net, dc.pi@nuaa.edu.cn
[2] Institute of Earthquake Science, China Earthquake Administration,
Beijing 100036, People's Republic of China
xiong.pan@gmail.com, shenxh@seis.ac.cn

Abstract. As one of the important methods on prediction, data mining plays a significant role specifically in the field of abnormal prediction to ensure security. Based on the remote sensing data of the sun-synchronous polar orbit EOS-AQUA satellite of USA, this paper proposes an abnormal pattern detection method with sequential pattern mining and matching. First of all, based on the selected observation area, abnormal sequential patterns are mined and frequent abnormal sequential patterns are formed. Then, seismic sequential pattern is generated, and the matching algorithm of earthquake is established. Finally, the accuracy rate and the false positive rate of prediction are worked out. All experiments are conducted with the remote sensing satellite from 2005 to 2014, and the experimental results are interesting. According to the carbon monoxide content, the accuracy rate is 65 % while the false positive rate is 15 % by using the data of 30 days before earthquake for prediction.

Keywords: Time series · Abnormal pattern · Data mining · Pattern matching

1 Introduction

It is widely concerned by international scholars to mine abnormal patterns by means of observing the changes of the contents of chemical gases on earth through satellite [1]. More importantly, as a perspective of abnormal mining and detection [2], studies on the changes of sequential pattern of chemical gases before the earthquake is a challenging topic worth further researching. Time series is one of the most typical data representations. Sequential pattern mining algorithm is mainly divided into two broad categories. One is based on the discovering association rules algorithm called Apriori, which was put forward by Agrawal R, Srikant R, et al. in 1995. And it includes not only AprioriAll, AprioriSome and DynamicSome algorithms, but also the derived Generalization algorithm for mining sequential patterns called Gsp, and SPADE [3] algorithm with vertical data format, etc. The other one is based on pattern growth proposed by Han, Pei, et al., including FreeSpan algorithm, PrefixSpan [4] algorithm,

Z. Huang et al. (Eds.): ICCCS 2015, LNCS 9483, pp. 480–490, 2015.
DOI: 10.1007/978-3-319-27051-7_41

which is quite different from the Apriori based algorithm and proved to be much more efficient.

In general, time series data has characteristics of high dimensions, and the choice of methods which represent the sequential pattern [5] is of great importance. The frequency domain representation maps time series to frequency domain space using the Discrete Fourier Transform (DFT), while Singular Value Decomposition [6] (SVD) represents the whole time series database integrally by dimensions reduction. Symbolic representation [7] is to map time series discretely to character string.

Studies on the emissions of chemical gas before the earthquake, such as, carbonic oxide (CO), methane (CH4), etc., are paid great attention to. Through the analysis of large area CO gas escaping from Qinghai Tibet Plateau on April 30, 2000, the Earth Observation System (EOS) reveals that there is anomalous layer structure in abnormal high CO content areas [8]. Supervised instances show that abnormal phenomenon before the earthquake exists objectively resulting from the increased emissions of greenhouse gases. According to the analysis of the 18 dimensions attributes of EOS-AQUA satellite data, it is shown through a large number of experiments that the CO content results of the abnormal sequence mining trend to be relatively good. Therefore, the experiments in the paper are based on the analysis CO content.

The rest of this paper is organized as follows. In Sect. 2, some related definitions are introduced. Section 3 is devoted to present the abnormal findings method upon sequence mining. The analysis of the experimental results is provided in Sect. 4. In final, the summary of this paper and future work are discussed in Sect. 5.

2 Related Definitions

Sequential pattern is viewed as a new method of earthquake prediction. For a more detailed understanding, some related definitions are given step by step as follows.

Definition 1 (Precursor time): We define precursor time as days before the day earthquake happened. So, the period of days is the precursor period of earthquake prediction. In order to find out the optimum prediction, precursor time of 30 days, 15 days and 7 days are adopted successively in this experiment.

Definition 2 (Precursor area): Precursor area is regarded as the region affected by seismic activities. For the sake of simplicity, the EOS-AQUA satellite data adopted in this experiment is partitioned into grids of 360 * 180. Besides, the distance between two points of the longitude is named level unit distance. By contrast, the distance between two points of the latitude is called vertical unit distance. Since there is no unified view on the division of precursor area, taking the length of level unit distance and level unit distance into consideration, we adopt two kinds of precursor area, namely 1 °* 1° and 2° * 2°, so as to find out the best one.

Definition 3 (Sequential pattern): If the support of sequence α, namely support(α) is no less than minsup, that is, α.sup ≥ minsup, sequence α is regarded as a sequential pattern in the sequence database. Moreover, sequential pattern with length of L is recorded as L-pattern.

Definition 4 (Sequence class): Sequences which is partly similar to each other are classified as a set, named sequence class. To be specific, Fig. 1 is the result of 10 seismic data sequential patterns, namely, the set of 10 sequential patterns. This sequence class is represented as $< S_1, S_2, S_3, S_4, S_5, S_6, S_7, S_8, S_9, S_{10} >$, where S_i stands for a mined frequent sequence of the data processed by symbolization.

a c c c c d
a a c c c c d d
c c c c c d
a c c c c c d d
a a c c c d d e
a a c c c c c d d e d
a a c c c c d
c c c d d
a c c c c c d d e
a c d d e d

Fig. 1. A collection of similar sequence-sequence class

Definition 5 (Sequence focus): The sequence, which gets the highest inclusive degree among all the sequences, is the focus of the sequences, referred to as sequence focus. Here, the inclusive degree of S_i is defined to be the ratio reflecting the degree how far the sequence S_i contains the other sequences in the same sequence class. Take sequence class in Fig. 1 for example, the sequence with the highest inclusive degree, which is 100 % here, is {a a c c c c c d d e d}, therefore we regard this sequence as the sequence focus of the sequence class.

Definition 6 (Difference set of sequential pattern): In view that seismic precursory data possibly contains non-seismic factors, we mine frequent sequences from both seismic data and non-seismic data. Then, difference set of sequential pattern is generated by subtracting the non-seismic sequence set from the seismic sequence set. That is, if one sequence from the frequent seismic sequence set occurs in the frequent non-seismic sequence set, the support of this sequence is subtracted and the sequence turns to be saved or abandoned depending on whether the subtracted support is no less than the initialized minimum support or not.

3 Sequential Pattern Mining and Matching Method

3.1 The Principle Diagram

In this paper, algorithms and experiments are proposed according to the following steps, with the flow chart depicted in Fig. 2.

Step 1. First of all, abnormal sequences are mined respectively from the processed seismic data and non-seismic data of the EOS-AQUA satellite. Meanwhile, frequent abnormal sequential patterns are generated accordingly, and marked as QuakeFreSet and NormalFreSet.

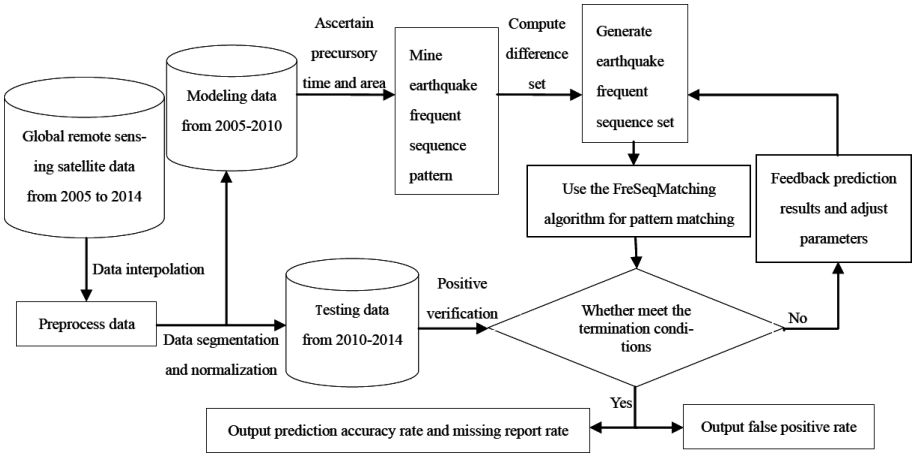

Fig. 2. The flow chart of abnormal findings method before earthquake

Step 2. In such a way that the frequent sequential patterns are generated, the specific sequential pattern before the earthquake is figured out. Moreover, sequence focuses meeting the defined conditions are located among the sequence class, after which the sets of sequence focus are formed as well as the matching algorithm.

Step 3. With the matching algorithm before the earthquake improved, the accuracy rate, the missing report rate and the false positive rate are computed to confirm the validity of this method.

3.2 Sequential Pattern Mining

In this experiment, PrefixSpan is adopted to mine frequent sequential patterns.

Algorithm PrefixSpan

Input : Sequence database S and the minimum support $min\_support$

Output : A set of complete sequence pattern

1：Read in the sequence database S and the minimum support threshould min_sup.

2：Set sequence length $K=1$ for the first time, and find out frequent sequence S with length of K from mapped database, where frequent sequence is no less than min_sup in the database.

3：By dividing the search space through S, respectively mine frequent sequnces, which obtain the $Prefix$ and sequence length of $K+1$. If the result of the mining empty, step 3 is turned to step 5.

4：Increase k to k+1, L founded in step 3 is assigned to S, and turn to step 2.

5：Record and output all the mined frequet sequence.

In addition, as a kind of depth first search algorithm, it maps the data to a smaller database recursively in the process of projection. On account of no need to generate candidate sequential patterns, the search space is shrunk as well as the scale of the projection database. Thereby, the efficiency of mining is enhanced to a great extent.

3.3 FreSeqMatching Algorithm

In this paper, we proposed a new matching algorithm named FreSeqMatching, which is responding to the matching degree of time series. For the sake of describing the matching algorithm clearly, a definition of matching function is provided as follows. To describe the matching algorithm clearly, related definitions are provided as follows.

$$match\_fun(F_i) = \begin{cases} 0, & isempty(LCS(\alpha, F_i)) = 1 \\ 1, & isempty(LCS(\alpha, F_i)) = 0 \end{cases} \quad (1)$$

Where, α represents a time series like <S_1, S_2, S_3 … S_n>, and F is the set of all the sequence focus, namely, {F_1, F_2, F_3 … F_i}. The function LCS(α, F_i) is used to get the longest common subsequence between sequence α and sequence focus F_i. If the longest common subsequence is empty, that is, isempty(LCS(α, F_i)) = 1, it means a failure match. Furthermore, the matching function is set to be 0, otherwise, to be 1.

The factors that influences the matching algorithm contain precursor time, precursor area, sequence support and data segment. In the case that the above parameters are set, matching degree can be further transformed to formula (2).

$$f\_\deg(\alpha) = \sum_{i=0}^{n} match\_fun(F_i) \div \sum_{i=0}^{n} F_i \quad (2)$$

Here, α and F_i play the same role as the above formula (1). By means of a large number of experiments, it turns out that when the matching degree belongs to [0.4, 0.7], the predicting results trends to be better.

$$f\_valid(F_i) = \begin{cases} 1, & f\_\deg(\alpha) \geq \sup Ratio \\ 0, & f\_\deg(\alpha) < \sup Ratio \end{cases} \quad (3)$$

It is indicated in Formula (3) that when the matching degree is no less than the defined support, the data is valid, namely, f_valid(F_i) = 1.

$$match\_num(F) = \sum_{i=0}^{n} f\_valid(F_i) \quad (4)$$

Formula (4) primarily aims to calculate the number of testing cases which is under certain condition, so as to work out both the accuracy rate and missing report rate.

The core concept of FreSeqMatching algorithm firstly is to positively verify seismic test set via the frequent sequence set, after which sequence matching degrees are figured out. Furthermore, seismic test data and non-seismic test data are matched by the mined frequent item sets.

Algorithm PreSeqMatching

Input : Frequent sequent set freqSeq, quake model data quakeModel, quake test data quakeTest, normal test data normalTest
Output : Accuracy rate matchRatio and missing report rate falseRatio
1: Read in frequent sequence set and data set of quakeModel, quakeTest, normalTest.
2: Initialize and set support supRatio.
3: Call function GetLFreq( ) to simplify freqSeq set.
4: Work out the model matching ratio of quakeModel set by the formula of modelMatchRatio = *MatchingDegree* (freqSeq, quakeModel).
5: If the result of modelMatchRatio meets the conditions, then turn to step 7.
6: If not, reset the support and supRatio turn to step 4.
7: For quakeTest set, calculate the matchRatio by the formula of matchRatio = *MatchingDegree* (freqSeq, quakeTest).
8: While for normalTest set, calculate the falseRatio through this formula falseRatio = *MatchingDegree* (freqSeq, normalTest).

Analysis:

(1) Step 1 and step 2 is for initialization. Step 3 aims at simplifying frequent sequence sets by GetLFreq function. With the purpose of backward verification through the modeling data, step 4 to 6 is in demand. What's more, step 7 is to calculate the prediction accuracy rate. Meanwhile, the false positive rate is worked out in step 8.
(2) The GetLFreq function above is used for simplifying frequent sequence sets.

Function GetLfreq

Input : Frequent sequence set freqSeq
Output : Simplified frequent sequences
1: Read in the number of freqSeq m.
2: Initialize the min_sup of freqent sequence.
3: For m frequent sequences, delete frequent sequences with support less than min_sup , update m.
4: For m frequent sequences, figure out the longest common sequence between every two sequences by comSeq = LCS(freqSeq[i],freqSeq[j]), and finally turn to step 8.
5: If comSeq is a intersection of the two sequences or empty, turn to step 4.
6: If comSeq = freqSeq[i] , then mark freqSeq[i] as flag , turn to step 4.
7: If comSeq = freqSeq[j], mark freqSeq[j] as flag, turn to step 4 as well.
8: Delete sequences marked with flag in step 4 to 7, and gain simplified frequent sequences.

(3) As an important function of FreSeqMatching algorithm, MatchingDegree function is repeatedly called in need, described as follows.

Function MatchingDegree

Input : Frequent sequence set freqSeq and quake test data quakeTest

Output : The accuracy rate matchRatio

1 : For each quakeTest[i] , for each simplified freqSeq[i], find out the longest common sequenct by the function LCS(quakeTest[i],freqSeq[j]).

2 : Calculate the value of match_fun(freqSeq[i]) through formula (2) above.

3 : For each earthquake case freqSeq[i] , figure out its matching degree based on formula (3), $\text{f\_deg}(\alpha)=\sum_{i=0}^{n} match\_fun\,(freqSeq[i]) \div \sum_{i=0}^{n} freqSeq[i]$.

4 : For each matching degree, calculate the value of f_valid(quakeTest[i]) by formula (4).

5 : According to given supRatio and formula (5), get the value of match_num(normalTest).

6 : Figure out the accuracy rate on sequence matching by matchRatio = match_num(quakeTest) $\div \sum_{i=0}^{n} quakeTest[i]$.

The LCS function in FreSeqMatching algorithm is a function with longest common subsequence and the content of sequence class and focus. Additionally, it is no longer described on this function in detail in this paper.

4 Experiments and Analysis

4.1 The Experimental Data Source

The experimental data covers EOS-AQUA satellite remote sensing data from the year 2005 to 2014, 217404000 data records in total. It contains 21 attributes, among which 18 attributes contribute to the seismic information.

Strong earthquake data with no less than 6.0 magnitudes is mainly adopted in this paper. The longitude of the selected earthquake area is from 73.5°E to 108.5°E, with the latitude from 20.5°N to 48.5°N. Mainly distributed in the western region in China, it covers the Qinghai-Tibet plateau seismic zone, etc. Moreover, it involves not only all or part of the Chinese provinces region, such as Tibet, Gansu, Yunnan, etc., and some part of neighbor countries, like Afghan, Pakistan, India, Bangladesh, Laos, etc.

4.2 Data Preprocessing

The remote sensing data of the EOS-AQUA satellite from 2005 to 2014 is divided into modeling data and test data. The classification of the satellite data is shown in Fig. 3.

As for the selection of test data, earthquakes with no less than 6 magnitudes are chosen from 2011 to 2014 as testing cases within the scope of 73.5°E to 108.5°E,

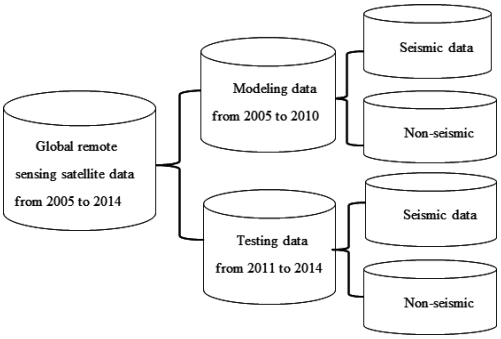

Fig. 3. Satellite data classification

20.5°N to 48.5°N. On account of the lack of enough earthquake cases, precursor area of 2°*2° is applied to obtain more earthquake samples in the experiment.

The main steps about data preprocessing are as follows.

(1) Data interpolation: among the remote sensing original data, outliers are represented by −9999, standing for the missing of the data. Nevertheless, it can be easily found that there is a certain amount of missing data. Therefore, data recovery is extremely necessary, namely, data interpolation. In this experiment, linear interpolation method is applied to take the place of the missing data appropriately.

(2) Data normalization: as a result of the influence of regional factors, remote sensing data are normalized in this paper. In view of the seasonal factors, the normalization in this experiment is corresponding to each month. That is, the mean values of all the historical data without earthquakes are computed in month, after which the percentage values divided by the average are figured out around 1. Hence, it can more effectively reflect the change trend of the data during the precursor time.

(3) Data segment: with the purpose of effectively representing the change trend of data, the linear segment method is applied on the basis of data normalization to turn into character representation. Consequently, it turns to be more convenient for mining sequential patterns. In order to gain better prediction results, different segments are adopted, such as 5, 7, 10 segments, to conduct experiments respectively.

4.3 Experimental Results and Analysis

Parameters Selection. The experiments are involved in a large number of parameters, with inclusive precursor time, precursor area, sequence support and the number of data segments, etc. Moreover, the selected parameters are briefly summarized in Table 1.

Table 1. Explanation of the selected parameters

Parameter list	Description of selected parameter
precursor time	Considering the best prediction effect, the experiment respectively selected 30 days, 15 days, 7 days before the earthquake as precursor time
precursor area	Taking the distance of level unit distance and vertical unit distance used in this experiment into account, two kinds of precursor area, 1°*1° and 2°*2, are adopted.
sequence support	In view of less earthquake cases, support of 0.3, 0.4, 0.5, 0.6 are set respectively.
data segment	On the basis of the normalization of data by month, segment of 5, 7, 10 are employed.

It is known from Table 1 that we have conducted 72 experiments to find out the better precursor time, precursor area, sequence support and data segments.

Analysis of Results. The prediction rate applied in the results is worked out as follows.

(1) SeismicData_CorrectRate, which is short for the correct rate of applying seismic data to predict earthquakes.

$$SeismicData\_CorrectRate = \frac{Tnum(SeismicDataTest\_True)}{Tnum(SeismicDataTest\_All)} \tag{5}$$

Where, Tnum(SeismicDataTest_True) refers to the number of correctly predicting earthquakes by seismic data, and Tnum(SeismicDataTest_All) points to the total number of the earthquake testing cases.

(2) SeismicData_FailureRate, standing for the failure rate of applying seismic data to predict earthquakes.

$$SeismicData\_FailureRate = 1 - SeismicData\_CorrectRate \tag{6}$$

(3) NormalData_FalseRate, which represents the false rate of using the normal data to predict earthquakes in this experiment

$$NormalData\_FalseRate = \frac{Tnum(NormalDataTest\_True)}{Tnum(NormalDataTest\_All)} \tag{7}$$

In formula (7), the number of correctly predicting earthquakes by non-seismic data is defined as Tnum(NormalDataTest_True), with Tnum(NormalDataTest_All) instead of the total number of the normal testing cases.

The accuracy rate, which comes from the carbon monoxide content (TOTCO_D) attribute with seismic data 30 days before the earthquake employed, is 65 % and the according missing report rate is 35 %. Meanwhile, non-seismic data 30 days before the earthquake is used to verify the experiments, and the false positive rate turns out to be

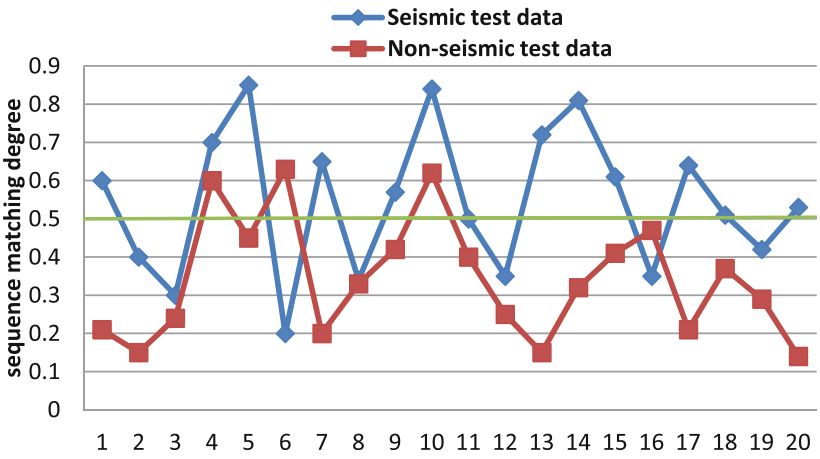

Fig. 4. Predicting results of the attribute of CO 30 days before earthquake

15 %. By contrast, the results are shown in Fig. 4, with X-axis to be the number of earthquake cases, and Y-axis to be the sequence matching support.

To explain Fig. 4 clearly, the sequence matching degree, which comes from the matching algorithm with the use of frequent patterns obtained from sequential pattern mining algorithm, reflects the similarity degrees between the testing cases and the mined earthquake frequent patterns. For seismic test data, it can be seen from the Fig. 4 that, when the matching support is set to be 0.5, the matching degree of NO.1 case is 0.6, greater than 0.5, so it is predicted to be seismic data. Whereas the matching degree of NO.2 case is 0.4, less than 0.5, it is conversely regarded as non-seismic data. As for non-seismic test data, NO.3 case is classified as non-seismic data, with matching degree of 0.24, obviously less than 0.5. Meanwhile, on account of the 0.63 matching degree, greater than 0.5, No.6 case is forecasted to be seismic data.

Hereby, there exist 13 cases of data with matching degree no less than 0.5 and 7 opposite cases among 20 cases of seismic data. Therefore, the accuracy rate is figured out to be 65 % based on Formula (5), with the missing report rate of 35 % on the basis of Formula (6). Besides, in 20 cases of non-seismic data, the number of cases with no less than 0.5 matching degree is 3, and the opposite is 17. Here comes the conclusive result that the false positive rate of prediction is 15 % in accordance with Formula (7).

5 Conclusions

It is an emerging direction of prediction to capture the exception rule by taking advantage of the technology of satellite for earth observation. From the perspective of time series, a method of abnormal pattern matching based on pattern mining is proposed in this paper, with the EOS-AQUA satellite data from 2005 to 2014. In final, after 72 times of experiments, it turns out that the predicting results of CO content is more satisfactory. Different from previous forecast model, it discovers abnormal

regular pattern of remote sensing data from a new point of view. As a consequence, effective abnormal patterns implied in the history are mined to realize the prediction preferably by pattern matching.

The prediction before the earthquake upon sequential pattern matching still remains several aspects to be improved as follows.

(1) If a better interpolation method is considered when replacing the invalid data, the actual missing value could be reflected more precisely, which makes the mined sequential pattern to be much more accurate to a certain extent.
(2) With time factor involved in discovered sequential patterns, a real-time prediction could gain more actual application value.

Acknowledgments. This paper is supported by National Natural Science Foundation of China (U1433116), High resolution seismic monitoring and emergency application (31-Y30B09-9001-13/15).

References

1. Alvan, H.V., Azad, F.H., Omar, H.B.: Chlorophyll concentration and surface temperature changes associated with earthquakes. Nat. Hazards **64**(1), 691–706 (2012)
2. Dong, X., Pi, D.C.: Novel method for hurricane trajectory prediction based on data mining. Natural Hazards Earth Syst. Sci. **13**, 3211–3220 (2013)
3. Zaki, M.J.: SPADE: an efficient algorithm for mining frequent sequences. Mach. Learn. **42**(1–2), 31–60 (2001)
4. Pei, J., Han, J., Mortazavi-Asl, B., et al.: Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: 2013 IEEE 29th International Conference on Data Engineering (ICDE), pp. 0215–0215. IEEE Computer Society (2001)
5. Bettaiah, V., Ranganath, H.S.: An analysis of time series representation methods: data mining applications perspective. In: Proceedings of the 2014 ACM Southeast Regional Conference, p. 16. ACM (2014)
6. Tong, X.H., Ye, Z., Xu, Y.S., et al.: A novel subpixel phase correlation method using singular value decomposition and unified random sample consensus. IEEE Trans. Geosci. Remote Sens. **53**, 4143–4156 (2015)
7. Baydogan, M.G., Runger, G.: Learning a symbolic representation for multivariate time series classification. Data Min. Knowl. Discov. **29**, 1–23 (2014)
8. Yao, Q., Qiang, Z., Wang, Y.: CO release from the Tibetan plateau before earthquakes and increasing temperature anomaly showing in thermal infrared images of satellite. Advances in earth science **20**(5), 505–510 (2005)