SEISMIC ANOMALY DETECTION USING SYMBOLIC REPRESENTATION METHODS

Vyron Christodoulou[1], **Yaxin Bi**[1], **George Wilkie**[1], **and Guoze Zhao**[2]

[1]*School of Computing and Mathematics, Ulster University, Jordanstown, Newtownabbey, Co. Antrim, BT37 0QB, U.K, e-mail: christodoulou-v@email.ulster.ac.uk, y.bi@ulster.ac.uk, fg.wilkie@ulster.ac.uk*

[2]*State Key Laboratory of Earthquake Dynamics, Institute of Geology, China Earthquake Administration Yard No.1, Hua Yan Li, Chaoyang District, Beijing, China, e-mail: zhaogz@ies.ac.cn*

ABSTRACT

In this work we investigate the use of symbolic representation methods for Anomaly Detection in electromagnetic sequential time series datasets. An issue that is often overlooked regarding the performance of symbolic representation methods with a sliding window in Anomaly Detection is the use of a quantitative accuracy measure. Until recently only visual representations have been used to show the efficiency of such algorithms. In this respect we propose a novel accuracy measure that takes into account the length of the sliding window and we present its utility. For the evaluation of the accuracy measure, HOT-SAX is used, a method that aggregates data points by use of sliding windows. A HOT-SAX variant, with the use of overlapping windows is also introduced, that achieves better results based on the newly defined accuracy measure. Both methods are evaluated on ten different benchmark datasets. Based on the empirical results we also evaluate them on Earth's electromagnetic data gathered by the SWARM satellites and ground-based sources around the epicenter of two seismic events in the Yunnan region of China.

Key words: seismic anomaly detection, symbolic representation, accuracy measure.

1. INTRODUCTION

Nowadays, the use of advanced data gathering methods routinely generates large volumes of data. Due to this trend, there are two approaches to the processing of large volume of data: (i) the increase of processing power and (ii) the reduction of the actual amount of data available. Within the domain of Anomaly Detection (AD), the importance lies in the fast indexing with minimum loss of information in order to identify any sequential anomalies and differentiate them from the normal cases. The use of an accuracy measure is considered standard practice for the evaluation and comparison of a method's performance. There are numerous different and useful methods to visualize and present how a method performs. Each one is more suitable for a particular task, as extensively researched in [14].

The existence of large volumes of data, formalized the need to transform these high volumes into a lower dimensional space. Such methods are extensively discussed in [8]. Within this large domain, there is a category of algorithms that uses a sliding window to aggregate and represent the original data points into a single symbol. This sliding window is usually user pre-defined. Such algorithms, which among them, HOT-SAX is the most ubiquitous [7], suffer from the problem that an accuracy measure, can not be applied. The issue arises because a measure that utilizes the length of the sliding window in its computation remains elusive. As such, studies around this area have been largely descriptive and qualitative. There has not been a clear quantitative measure on the performance of these methods.

A qualitative approach in the presentation of the results is only useful to understand how an algorithm processes and produces its outcome. Most improvement in numeracy based research comes from a quantitative point of view. Therefore it is imperative to define an accuracy measure that takes into account the sliding window component of such algorithms. Defining the accuracy, paves the way for research to move forward. Moreover, this assists in understanding an algorithms' drawbacks, such as the tendency to predict false positives or miss genuine anomalies.

A fundamental contribution of this work is therefore, the development and definition of the accuracy measure. Its performance is evaluated in detail by evaluating the original HOT-SAX in ten benchmark datasets. The results give us an understanding on how the algorithm works, pinpointing issues that arise and what more can be changed to improve it. Based on the accuracy and our understanding, an improvement of the base HOT-SAX is proposed. This variant introduces the use of an overlapping sliding window. A sliding window is synonymous with Fast Fourier Transformation (FFT). An overlapping window is known to offer additional information in time series processing. Anomalies or events, otherwise missed from a non-overlapping window can be captured by using an overlap. Therefore although the idea is not new, it is a well established fact in the community [15]. The accuracy measure is then used to compare both algorithms. With the definition of the accuracy measure, the problem becomes trivial as it provides quantitative results for comparison. This category of algorithms can then be fine-tuned and improved based on the accuracy measure.

The rest of the paper is organized as follows: In Section II, the background of the work is set and relevant research is presented. In Section III, the methodology is discussed and the accuracy measure is defined. Following that, the original HOT-SAX algorithm and the proposed variant are presented. In Section IV, the empirical results are shown where both algorithms are evaluated against ten benchmark datasets. Based on how their best performance they are tested in real ground-based and satellite data. Lastly, in Section V, the discussion and possible future developments are proposed.

2. BACKGROUND AND RELATED WORK

Similarity search and AD in particular suffer from the exponential growth of the search space caused by the high volume of data, that makes efficient data processing unattainable. In order to circumvent the curse of dimensionality problem, the efficient processing of time series data requires a lower dimensional approximation.

The focus of this work is on data adaptive methods, specifically on the symbolic representation of time series data. Some of the earlier methods on symbolic string representation are SDA [2]. SDA requires the user to divide a range from one point to another and the algorithm computes the changes between these points. IMPACTS[3] yet another method, incorporates a similar technique but it computes the change ratio between one point and the next in order to discretize them into equal sized bins and symbolically represent them. Since then, research has come a long way and most recently symbolic representation has become popular for use with high volumes of data. A good symbolic representation of time series data offers high discretization, better scalability, high readability and can benefit from other well-researched fields that utilize similar methods such as text mining, bioinformatics and chemoinformatics. For that reason one of the first novel methods that was focused and designed for AD was first introduced in [16]. The authors propose the idea of a negative selection mechanism based on the human immune system. Self and non-self patterns, match the normal and anomalous sequences for achieving the AD.

The struggle for simplicity and less parameter configuration led to the development of one of the most popular symbolic representations. HOT-SAX, that is going to be reviewed later, uses the Piecewise Aggregate Approximation (PAA) as its core component. A multitude of variants have been developed which include among others: HOT-aSAX [17] that uses a k-means to decide the number of breakpoints for the cartesian space to be segmented, 1d-SAX [18] that together with the average provides an additional level of resolution by computing the linear regression of each subsequence or Symbolic Fourier Approximation [19] that uses Fourier coefficients from each subsequence for the symbolic approximation.

All of the above methods share one common component: they all make use of a user selected window to extract and combine, albeit in a different way, all the subsequences involved in the time series. Determining a sliding window requires prior-knowledge of the time series data. Within the AD domain, usually the focus has been on providing an anomaly score [13], or when using a sliding window a visual interpretation or visual mining [6], [7], [11] for the evaluation and comparison with other methods. Most comparisons consider the pruning power of each algorithm for dimensionality reduction, the number of calls to the distance function for a time efficient AD or the mean wall clock time for computing how long takes for a process to finish [9]. However, to our knowledge, until now there has not been a comparison using the accuracy for AD and thus a definition of a precise accuracy measure that combines the window length and the F1-score to make a unified accuracy measure is required.

Different assessment measures have been proposed depending on the problem under scrutiny [1]. The assessment methods yield different results based on their optimization. Others underestimate certain classification capabilities, i.e by producing less false positives, whereas other overestimate them. Ultimately, what an accuracy measure should achieve is to indicate whether a method can generalize well in unknown data within the same domain. The F1-score, in Information Retrieval, was developed for a common and more reliable measure than the accuracy [1]. It can be interpreted as a weighted average of the precision and recall. From Eq.1 we can see that there is an inherent bias towards the anomalies, since the F1-score does not measure the True Negatives. For that reason there was a need to define a more balanced measure.

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{1}$$

Drawing from the field of bioinformatics, there are multiple measures used to calculate the True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN) and their relations. These include the sensitivity, specificity, recall or ROC and AUC curves. Nevertheless, an important contribution was the Matthew's correlation coefficient (MCC) [10]. Its importance lies in the fact that it gives balance to all the parameters in a classification task. However, it can be high in cases where there are very few FP and at the same time very few TP cases. For reference, MCC is given by the formula:

$$MCC = \frac{TP{\times}TN - FP{\times}FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

The MCC therefore becomes an non balanced measure in AD. Since any performance measure is based on the TP, FP, TN and FN and their relation if any of these values is overlooked it means loss of information. However, when there is an interest in one class, as in AD, the F1-Score becomes the appropriate measure to use since it provides a weighted average of the anomalies. As a consequence the amount of TN is of no relevance to the proposed measure. In order to address the issue of not taking into account the length of the sliding window, we propose a measure for binary classification that unifies both the F1-Score and the window length for AD in data adaptive aggregation methods.

3. METHODOLOGY

3.1. Definitions

The definitions that are used throughout the work are provided here:

Time Series: A time series $C = c_1, c_2...c_m$ of length m is a set of real ordered values that follow a m-order.

Subsequence: From a time series C of length m a subsequence S of C is a series of length $n \leq m$ of adjacent position from C, that fulfils the equation, $S = c_i, ...c_{i+n-1}$ for $1 \leq i \leq m - n + 1$

Sliding Window: A sliding window, W, is a user defined subsequence of length n, which can be used to extract all given subsequences S_i if we slide it across a time series C of length m.

Overlap: An overlap or offset L is a user-defined parameter of size p, that covers a percentile of length n*p of the previous sliding window and extracts all possible subsequences from a time series, C, of size m. The final number of the extracted subsequences is $u = (m - n)/(n - p) + 1$, where n is the length of the subsequence, p is the overlapping size and m the length of the time series.

Anomaly: A subsequence of length n, that begins from the position p and is a part of the time series C. It is said to be an anomaly when it has the largest distance from its nearest non-self match.

The bruteforce algorithm that is used for the detection of the anomalous subsequences uses the euclidean distance,

Euclidean Distance:

Given two time series A and B of length m, the euclidean distance between them:

$$Dist(A, B) = \sqrt{\sum_{l=1}^{m} (q_l - c_l)^2} \quad (2)$$

3.2. Data and Preprocessing

The ten benchmark data used were downloaded from the physionet website[1] and include more than one anomalies from a variety of physiological signals. Out of them, two databases were selected: (i) the chf database, that contains congestive heart failure data and (ii) the mit database, that contains arrhythmia related data. Each dataset contains two channel ECG recordings. The length and the channel of the benchmark time series datasets used are denoted by the subscript and the superscript respectively, $dataset_l^c$. Finally, the annotated and exact anomalous locations were marked by cardiologists.

The real satellite datasets used were acquired by the ESA website[2]. They consist of 72 consecutive days of observations that three identical SWARM satellites gather at a height of 450km for SWARM A and C and at 530km for SWARM B. The observations range from 21^{st} November 2014 to 31^{st} January 2015. The data are in the Common Data Format (CDF) and consist of 22 different fields with 86400 values, one for each second of a day. The relevant data are extracted from the Vector Field Magnetometer (VFM) that cover our study area and measure Earth's electromagnetic field intensity. The area of interest is around the vicinity of the occurred seismic events and the central grid point is focused around the coordinates [23.358^o N, 100.5333^o E], [23.336^o 100.474^oE] where the seismic events, of a scale $M_g = 5.6$ occurred on the 5^{th} December at 10:20:01 UTC and 6^{th} December 18:43:46 UTC respectively.

The terrestrial data are from the X and Y vectors of Earth's electromagnetic field gathered by a Control Source Extremely Low Frequency (CSELF) observatory in Jinggu, at coordinates N [23.30^oN, 100.44^oE]. The data were averaged to form an 1-D vector of 72 observations from the same time range as the satellite data. The area of the events is shown in Fig. 1 situated in the Yunnan region of China.

Earth's electromagnetic field, B is described by three orthogonal components: X, the northerly, Y the easterly and Z the vertical intensity. The final intensity vector can be calculated from the orthogonal components using Eq.3 [20]. The process followed for the data preprocessing is:

- Extract the values from the coordinates that belong to the region of study.
- Convert the X, Y, Z orthogonal vectors to a single vector.

$$|\vec{B}| = \sqrt{b_X^2 + b_Y^2 + b_Z^2} \quad (3)$$

- Subdivide the region of study by using the grid and create nine different vectors.
- Take into account all the observations from the three different satellites and form a single vector for the 72 days.
- Take the mean of each day for each vector.

[1] https://physionet.org/cgi-bin/atm/ATM

[2] https://earth.esa.int/web/guest/swarm/data-access

- Create the 72 data point 1-D vector for each of the nine grid elements. Each data point corresponds to a single day.

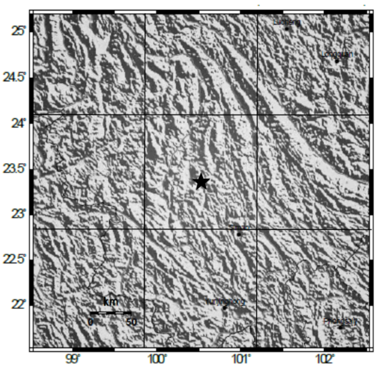


Figure 1: The study area in the Yunnan region of China

It is worth noting that the preprocessing has some caveats. Each satellite has 3 to 4 days revisit time. This makes apparent the issue of the level of resolution that has to be applied to the study. A very small study area might not contain enough observations since the satellite revisit location might be different. The very few data make this idea unworkable and it forces the need to opt for a larger area. The chosen area is 2000km x 2000km from [13 -33 , 90 E- 110 E]. Each grid point is 666.6km. The choice was made because SWARM A and SWARM C that fly in parallel, have a distance of 150km. Taking roughly four times their distance ensures more observations within a grid point. Nevertheless, the problem is not solved in its entirety and missing values were interpolated by calculating the mean of the neighbouring points.

3.3. A Review of HOT-SAX

One of the most well known methods for time series symbolic approximation is HOT-SAX. HOT-SAX utilizes a piecewise aggregate approximation (PAA) and assumes that the PAA follows a Gaussian distribution. It transforms a numerical time series into a symbolic approximation of the original of a finite alphabet cardinality. SAX works in three simple steps 2:

- Divide time series C into subsequences of length n.
- Calculate the mean of each subsequence S.
- Discretize each value into equidistant bins following the Gaussian distribution based on an alphabet of cardinality Y. The break up points can be calculated by a statistical look up table.

3.4. The HOT-SAX variant

HOT-SAX is very dependent on the user selected window length. Every aggregation method is dependent on how much information is preserved. If the length, n of the sliding window, W, is smaller than the anomaly, the aggregation step can miss or smooth the anomaly. Similarly, if the window is several scales of magnitude larger than the length of the anomaly the anomaly can also be missed. Although prior knowledge of the dataset is evidently required, the proposed addition of an overlap tries to alleviate this over-dependency on the user selected length of the sliding window.

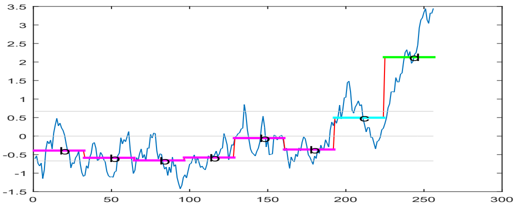


Figure 2: A visual representation of the three SAX steps

The overlapping window works by sliding the window over the time series but with the inclusion of an overlap. Effectively, this helps capture anomalies easier because once an anomaly is present in the time series, the overlapping window carries the anomaly forward to the next aggregation. Therefore because of the overlap the calculation leaves a residue of the anomalous data points onto the next PAA, as seen in Fig.3.

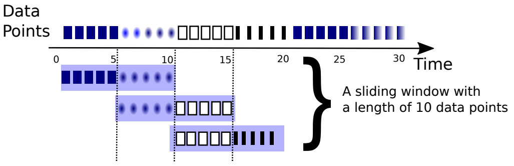


Figure 3: A sliding window with 50% overlap.

This addition helps focus on the granularity of the algorithm and offers additional resolution. The algorithm can identify anomalies even if the window length is smaller than the anomaly. The improvement of the variant algorithm over the base algorithm can be assessed by the introduction of the accuracy measure in the next section. With the accuracy measure, if an anomalous data point falls within the user selected window it is considered as a TP. If the anomaly lies outside the window it is considered to be a FP. If an identified anomaly falls within the window but it is not a true anomaly it is a FN.

3.5. The Accuracy Metric

In order to define how different a prediction is from the true value the definition of an error has to be given. Therefore, the error from the predicted versus the true anomalous location, is given by the following equation:

$$e = t_w - p_w \tag{4}$$

where, t_w is the true anomalous location and p_w is what the algorithm predicted.

The bruteforce algorithm pinpoints the exact locations of the predicted anomalies. This is considered to be the midpoint of the sliding window, *W*. Given the Equation 4 the error, meaning how far the true anomalous location from the predicted by the algorithm is, becomes a factor of the window and has a range of $p_w - W/2$ to $p_w + W/2$ based on the predicted location.

Now, let all the anomalous locations identified by the bruteforce algorithm to belong to the set P. Therefore,

$$P = \{p_1, p_2, ...p_s\} \tag{5}$$

Let T be a set of the locations that the experts have identified as true. Thus, there is:

$$T = \{t_1, t_2, ...t_t\} \tag{6}$$

In all, the locations that interest us are formed by Eq.7. The correctly predicted anomalous locations belong to a new set N which is formed by the intersection of P and T and represents the TP.

$$N = T \cap P \tag{7}$$

More specifically, during the AD a true anomalous location might or might not fall within the range that is based on the predicted location and the window. It is reasonable to say that when both locations fall within this range, their intersection will cause them to form a set on their own. Therefore the final locations' set will be:

$$N = \begin{cases} N \supseteq \{T, P\}, & \text{if e} \leq \text{W} \\ N = P, & \text{otherwise} \end{cases} \tag{8}$$

Consequently, the true locations that fall within the range of the predicted anomalous locations form the elements of the strict superset *N*, *k*. If more than one anomalies fall within the same range, this case is treated by the algorithm as a unique entity or a single anomaly. As a result, the first case in Eq.8 is considered as a TP and the second case as a FP. In the accuracy measure the TP values, the positive class, that form the set N are used as the averaging factor.

After the definition of the basic factors, the definition of the accuracy becomes:

$$A = \sum_{1}^{k} \frac{F_1 - (Z_k \times F_1)}{N} \tag{9}$$

where,

$$Z_k = \frac{|t_k - p_k|}{\frac{W}{2}}$$

It is known that precision, recall and F1-Score ignore the TN values and can have a positive bias. Nevertheless, because of that and since our interest lies specifically in the positive class, meaning the class or set of correctly or incorrectly predicted anomalous cases this set is used as the main factor in the definition of our accuracy measure. As already mentioned, the F1-Score is the harmonic mean and an indication on how the precision and recall work. In all, the defined accuracy gives the average error rate for the correctly predicted values based on a factor of the window length.

4. EMPIRICAL RESULTS

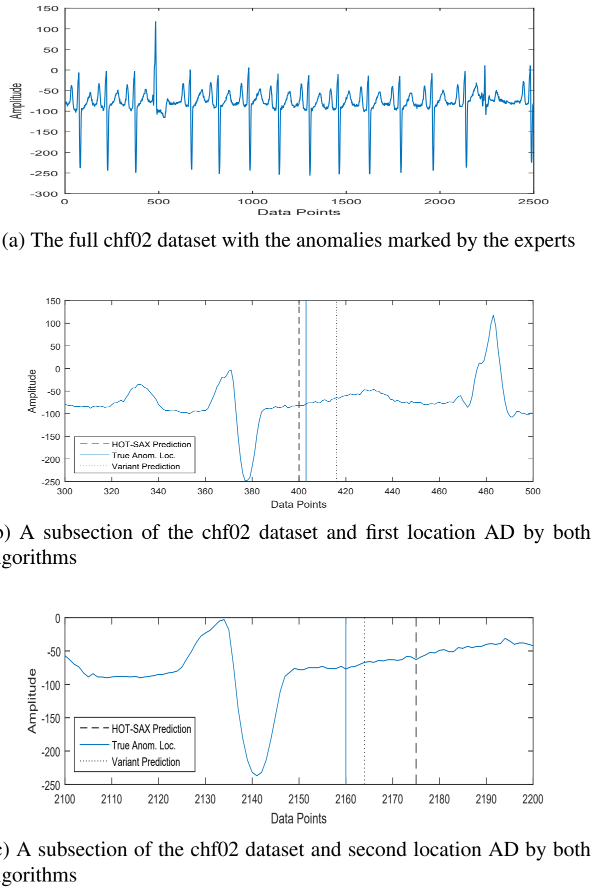
The results for the HOT-SAX variant show an improvement of 30.25% on average from the base algorithm. The use of overlapping windows for the symbolic representation of the time series makes the anomalies more "visible" and impactful, as shown by the accuracy measure.

Table 1: Accuracy for HOT-SAX and HOT-SAX with overlap

Dataset	Overlap (%)	Parameters	Bruteforce	Accuracy (%)
$chf02_1^{2500}$	0	P_{100}^{10}	3	62.6
	0.5	P_{100}^{15}	6	**69.6**
$chf02_2^{2500}$	0	P_{50}^{15}	2	37
	0.6	P_{100}^{15}	2	**56**
$mitdb108_1^{15000}$	0	P_{100}^{15}	5	30.52
	0.6	P_{100}^{15}	6	**55.95**
$mitdb108_2^{15000}$	0	P_{100}^{15}	5	45.8
	0.4	P_{100}^{10}	7	**61.7**
$mitdb101_1^{3600}$	0	P_{400}^{15}	3	27.5
	0.5	P_{400}^{15}	3	**71**
$mitdb101_2^{3600}$	0	P_{100}^{15}	4	28
	0.4	P_{400}^{15}	5	**82.5**
$mitdb100_1^{15000}$	0	P_{1000}^{15}	4	41.7
	0.2	P_{1500}^{15}	15	**69**
$mitdb100_2^{15000}$	0	P_{1000}^{15}	15	41.7
	0.4	P_{400}^{15}	8	**62.93**
$chf12_1^{3600}$	0	P_{250}^{10}	4	36.8
	0.4	P_{250}^{15}	4	**71.2**
$chf12_2^{3600}$	0	P_{250}^{10}	4	36.8
	0.4	P_{1000}^{15}	4	**60.98**

The perennial issue of this kind of algorithms is the a-priori knowledge of the datasets that is required for their configuration. However, in real life this rarely happens. The single most important parameter is the length of the subsequence to consider to encode it into a symbol. Nonetheless, given the intuitive nature of the parameters, one can have an understanding on how the algorithm works and what are its limitations.

One of the most important limitations is the length versus the alphabet cardinality we need to consider. If the cardinality is very small compared to the length of the window, the algorithm produces unattainable results. Similarly, an arbitrarily high cardinality number can be regarded as simply noise. This occurs because a very high segmentation of the cartesian space provides very high resolution and the PAA results become meaningless.



(a) The full chf02 dataset with the anomalies marked by the experts

(b) A subsection of the chf02 dataset and first location AD by both algorithms

(c) A subsection of the chf02 dataset and second location AD by both algorithms

Figure 4: AD process for both algorithms

Fig. 4 shows the AD in the dataset $chf02_{1}^{2500}$ for both algorithms. The actual anomalies are at points 403 and 2160. The base algorithm identifies an anomaly at points 400 and 2175 respectively. The variant algorithm in 412 and 2164. In this case the HOT-SAX variant performs better, because although the base did better in the first anomaly, in the second case the detection of the anomaly is further away than the actual location. Both algorithms also had a FP location lying outside the boundaries of the more increased resolution versions as seen in the figure, something that is also visible in their detection accuracy in Table 1. The method's user selected inputs, the alphabet cardinality and the anomaly length are denoted as the superscript and subscript respectively, P_{a}^{r}. For all other datasets a similar detection process is followed and we present the results for all the different benchmark datasets Table 1. All benchmark datasets had more than one anomalies present. Another aspect is how the accuracy measure performs. A factor of the F1 score is calculated each time for each of the anomalous detected points. The accuracy is also affected by the length of the bruteforce window during the AD process. A saturation point can be reached when a window comparable to the length of the symbolic representation is selected. The same saturation point can be reached much faster if a very small alphabet cardinality is selected, as previously explained. As a matter of fact, the AD does not work after the saturation point. A graphical representation of the issue can be seen in Fig. 5.

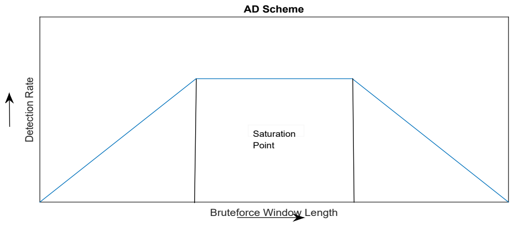


Figure 5: The Saturation point of the AD

4.1. SWARM and CSELF Datasets

Equipped with the knowledge from the benchmark datasets and knowing the limitations of the algorithms, we use the same principles to the real datasets. From the empirical results we follow the same rules accordingly. The window length is set as 3, a larger window compresses too much the original signal, and loses a lot of information. For reference, the length of the final sequence from the base algorithm is 24 symbols since the window is 3 days.The cardinality used is 15, the same that achieved the best results in the benchmark datasets. The bruteforce window was selected as 5 for the base algorithm and for the variant, both based on the benchmark datasets performance and taking into account the scale of the real data after the symbolic representation. Lastly, the overlap selected was 40%, again based on the empirical results.

The results from all the grid-points are presented in Fig.6. Although it is not possible to verify the findings, the results are in line with other studies in the area of earthquake precursory anomalies that identify anomalies within a similar time frame [4],[5]. Other causes of magnetic interference or noise might be the cause of anomalies. Nonetheless, the reliability and the results of both data can be of use. By identifying the key differences between terrestrial and satellite data the key question becomes our ability to understand anomaly patterns.

5. DISCUSSION AND FUTURE WORK

The evaluation of both algorithms on the benchmark datasets from the accuracy measure helped us interpret the experimental results in a new way. The understanding of how to fine-tune state of the art algorithms based on

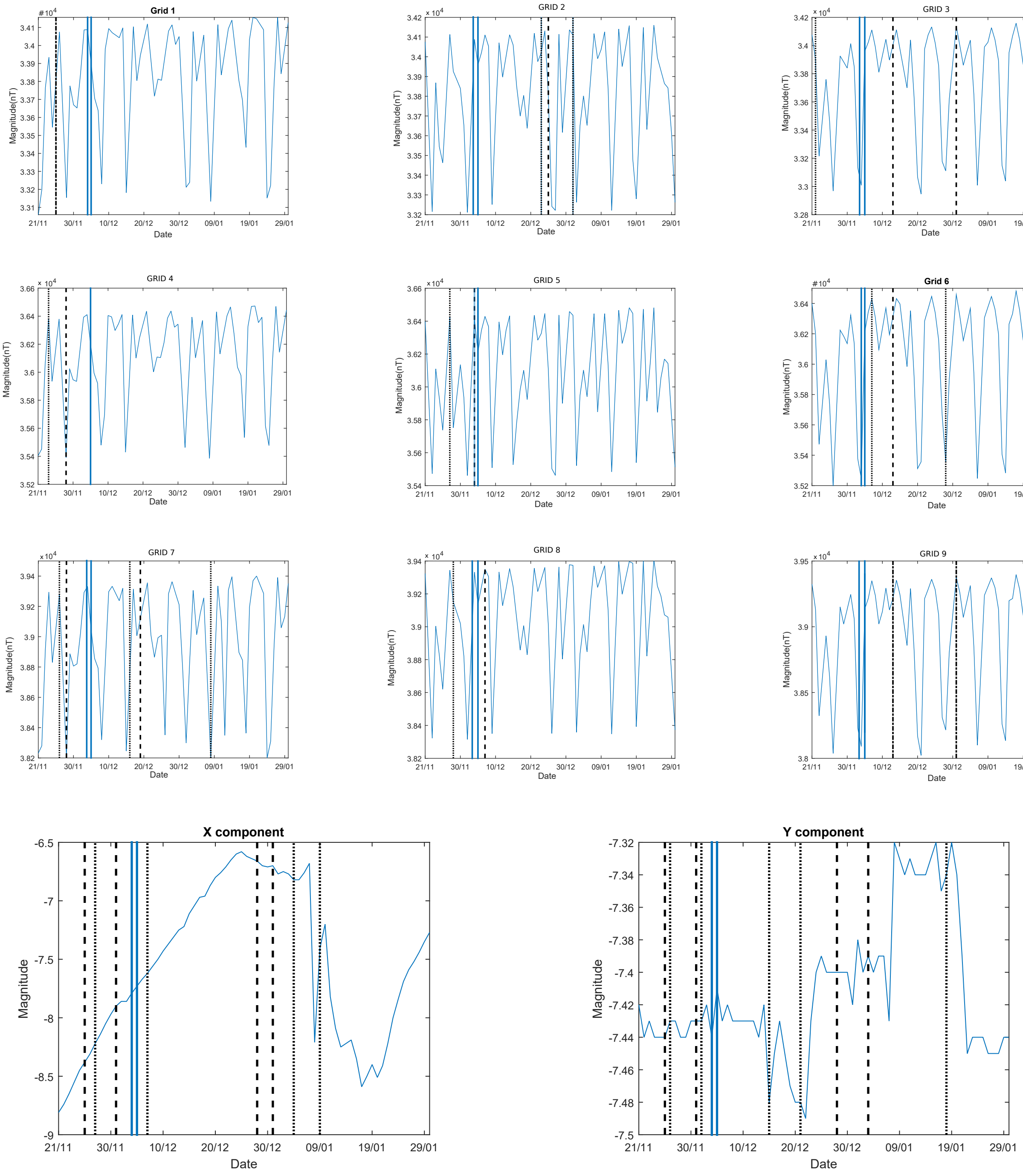


Figure 6: Anomaly Detection for both algorithms applied to the real dataset, *Top to Bottom*: Grid point 1 to 9 and the 2 CSELF datasets. *Solid vertical line*: Occurred earthquakes, *Dotted vertical line*: HOT-SAX Variant algorithm prediction, *Dashed vertical line*: HOT-SAX prediction

the proposed accuracy measure, further expands our ability to work and expand them to real datasets and novel research. It is important to know how exactly an algorithm fares when one parameter is changed and how much it affects the other parameters and the final accuracy.

Symbolic representations prove to be accurate and with the combination of their time efficiency they can be considered for online and real time AD. Ultimately the resolution versus the loss of information has to be weighted in the final choice of representation. Their application in the real datasets, showcases their advantages and reveals some interesting results. Based on the comparison of the terrestrial and satellite sources we can get an understanding both on (i) the reliability of data and (ii) the resolution required for further studies. The results follow a large research that also is able to detect anomalies before seismic events but they must be put under further scrutiny. Symbolic representation has a lot to offer in processing large datasets such as the ones used in this study from satellites and offers new possibilities when used for AD. The key question on the identification of anomalies as precursors to earthquakes is still open and the results suggest the need for further investigation.

ACKNOWLEDGMENTS

This work is supported by the Dragon 3 program, a collaboration between the European Space Agency and the Ministry of Science and Technology of China. The authors would also like to acknowledge some additional support received from the project funded by the National Natural Science Foundation of China (Grant No. 41374077) and the European Space Agency for making SWARM data available for our research.

REFERENCES

[1] C. J. Van Rijsbergen. 1979. Information Retrieval (2nd ed.). Butterworth-Heinemann, Newton, MA, USA.

[2] Andr-Jnsson, Henrik, and Dushan Z. Badal. ”Using signature files for querying time-series data.” Principles of Data Mining and Knowledge Discovery. Springer Berlin Heidelberg, 1997. 211-220.

[3] Huang, Yun-Wu, and Philip S. Yu. ”Adaptive query processing for time-series data.” Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 1999.

[4] Akhoondzadeh, M. (2013). Genetic algorithm for TEC seismo-ionospheric anomalies detection around the time of the Solomon (Mw= 8.0) earthquake of 06 February 2013. Adv. Space Res., 52(4), 581-59

[5] Kong, X., Bi, Y., Glass, D. H. (2015). Detecting Seismic Anomalies in Outgoing Long-Wave Radiation Data. Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of, 8(2), 649-660

[6] Lin, Jessica, et al. ”VizTree: a tool for visually mining and monitoring massive time series databases.” Proceedings of the Thirtieth international conference on Very large data bases-Volume 30. VLDB Endowment, 2004.

[7] Keogh, Eamonn, Jessica Lin, and Ada Fu. ”Hot sax: Efficiently finding the most unusual time series subsequence.” Data mining, fifth IEEE international conference on. IEEE, 2005.

[8] Zimek, Arthur, Erich Schubert, and HansPeter Kriegel. ”A survey on unsupervised outlier detection in highdimensional numerical data.” Statistical Analysis and Data Mining 5.5 (2012): 363-387.

[9] Lin, Jessica, et al. ”A symbolic representation of time series, with implications for streaming algorithms.” Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. ACM, 2003.

[10] Matthews, Brian W. ”Comparison of the predicted and observed secondary structure of T4 phage lysozyme.” Biochimica et Biophysica Acta (BBA)-Protein Structure 405.2 (1975): 442-451.

[11] Rebbapragada, Umaa, et al. ”Finding anomalous periodic time series.” Machine learning 74.3 (2009): 281-313.

[12] Baldi, Pierre, et al. ”Assessing the accuracy of prediction algorithms for classification: an overview.” Bioinformatics 16.5 (2000): 412-424.

[13] Izakian, Hesam, and Ajith Abraham. ”Fuzzy C-means and fuzzy swarm for fuzzy clustering problem.” Expert Systems with Applications 38.3 (2011): 1835-1838.

[14] Gunawardana, Asela, and Guy Shani. ”A survey of accuracy evaluation measures of recommendation tasks.” The Journal of Machine Learning Research 10 (2009): 2935-2962.

[15] Bastiaans, Martin J. ”On the sliding-window representation in digital signal processing.” Acoustics, Speech and Signal Processing, IEEE Transactions on 33.4 (1985): 868-873.

[16] Dasgupta, Dipankar, and Stephanie Forrest. ”An Anomaly Detection Algorithm Inspired by the Immune System.” Artificial immune systems and their applications. Springer Berlin Heidelberg, 1999. 262-277.

[17] Pham, Ninh D., Quang Loc Le, and Tran Khanh Dang. ”HOT aSAX: A novel adaptive symbolic representation for time series discords discovery.” Intelligent Information and Database Systems. Springer Berlin Heidelberg, 2010. 113-121.

[18] Malinowski, Simon, et al. ”1d-sax: A novel symbolic representation for time series.” Advances in Intelligent Data Analysis XII. Springer Berlin Heidelberg, 2013. 273-284.

[19] Schfer, Patrick, and Mikael Hgqvist. ”SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets.” Proceedings of the 15th International Conference on Extending Database Technology. ACM, 2012.

[20] Campbell, Wallace H. Introduction to geomagnetic fields. Cambridge University Press, 2003.