For review for presentation in **CDC ’09**, Shanghai, China

Optimization of Time-series Data Partitioning for Anomaly Detection

Xin Jin xuj103@psu.edu
Soumik Sarkar szs200@psu.edu
Kushal Mukherjee kum162@psu.edu
Asok Ray axr2@psu.edu

Department of Mechanical Engineering
The Pennsylvania State University
University Park, PA 16802, USA

Keywords: *Optimization; Partitioning; Time series analysis; Anomaly detection; Maximum migration*

***Abstract*— The concepts of symbolic dynamics and data set partitioning have been used for feature extraction and anomaly detection in time series data. Although modeling of state machines from symbol sequences has been widely reported, similar efforts have not been expended to investigate partitioning of time series data to optimally generate symbol sequences for anomaly detection. This paper addresses this issue and proposes a partitioning method based on maximum migration of data points across cell boundaries. Various aspects of the proposed partitioning tool, such as adaptiveness of alphabet size selection, noise mitigation, and robustness to spurious disturbances, are discussed. Experimental results on laboratory apparatuses of electronic circuits and electric motors show that maximum-migration partitioning yields significant improvement over existing partitioning methods (e.g., maximum entropy partitioning) for the purpose of anomaly detection.**

1. Introduction

Early detection of anomalies (i.e., deviations from nominal behavior) in human-engineered complex dynamical systems is essential for prevention of catastrophic failures, enhancement of performance and survivability. In general, the success of data-driven anomaly detection techniques depends on the quality of feature extraction from sensor time-series data. To this end, several tools of feature extraction tools, such as principal component analysis (PCA) [1], independent component analysis (ICA) [2], kernel PCA [3], and semidefinite embedding [4], have been reported in literature. Symbolic Dynamic Filtering (SDF) [5] is a data-driven tool of anomaly detection, which is built upon the concepts of symbolic dynamics. SDF also serves as a feature extraction method via partitioning of time-series data to generate symbol sequences.

Although modeling of state machines from symbol sequences has been widely reported, similar efforts have not been expended to investigate partitioning of time series data to optimally generate symbol sequences for anomaly detection. Stauer et al. [6] reported comparison of maximum entropy partitioning and uniform partitioning; it was concluded that maximum entropy partitioning is a better tool for change detection in symbol sequences than uniform partitioning. Symbolic false nearest neighbor partitioning (SFNNP) optimizes a generating partition by avoiding topological degeneracy. However, a shortcoming of SFNNP is that it may become extremely computation intensive if the dimension of the phase space of the underlying dynamical system is large. Furthermore, if the time series data become noise-corrupted, the symbolic false neighbors rapidly grow in number and may erroneously require a large number of symbols to capture pertinent information on the system dynamics [7]. The wavelet transform largely alleviates the above shortcoming and is particularly effective with noisy data for large-dimensional dynamical systems [8]. Maximum entropy partitioning was utilized to convert the wavelet space partitioning (WSP) data to a sequence of symbols. Although WSP is significantly computationally faster than SFNNP and is suitable for real-time applications, WSP too has several shortcomings including: requirements of good understanding of signal characteristics for selection of the wavelet basis, identification of appropriate scales, and conversion of the two-dimensional scale-shift domain into a single dimension. Subbu and Ray [7] introduced Hilbert-transform-based analytic signal space partitioning (ASSP) as an alternative to WSP. Sarkar et. al [9] generalized ASSP for symbolic analysis of noisy signals. Nevertheless, these partitioning techniques primarily attempts to provide an accurate symbolic representation of the underlying dynamical system under a given quasistationary condition. Therefore, a partitioning that focuses on statistical changes of time-series data as the dynamical system moves from one quasistationary condition to another may prove to be more useful for the purpose of anomaly detection.

SDF can be interpreted as a tool for compressing and transferring dynamical system information from the space of time-series data to the space of patterns via data partitioning and state machine construction. Properties and variations of transformation from symbol space to pattern space have been thoroughly studied in mathematics, computer science and especially data mining literature. Although quite a few variations of partitioning techniques are reported in the physics and signal processing literature, only a few of them address the issue of anomaly detection in particular. This paper proposes a partitioning method based on maximized migration of time series data points across cell boundaries during evolution of the underlying dynamical system. To this end, a framework is presented toward optimization of this partitioning scheme for anomaly detection and investigates its major features, namely, robustness of extracted information from symbol sequences and enhancement of computation efficiency. The paper is organized into four sections including the present section. The partitioning scheme is described in an optimization framework in Section 2 along with its key features. Section 3 validates the proposed concepts on an active electronic circuit apparatus that implements a second order non-autonomous forced Duffing equation [10]. Section 4 summarizes the paper and makes major conclusions along with recommendations for future research.

2. Optimization of the Partitioning Scheme

Recent literature [5] [8] has explored the concepts of symbolic dynamics and data set partitioning to develop a computationally efficient tool called Symbolic Dynamic Filtering (SDF) for anomaly detection in complex dynamical systems. Although the SDF methodology is reported in recent literature, a brief outline of the procedure is succinctly presented here for completeness of the paper.

Anomaly detection from time series data is posed as a two-scale problem. The *fast scale* is related to the response time of the process dynamics. Over the span of data acquisition, dynamic behavior of the system is assumed to remain invariant, i.e., the process is quasi-stationary at the fast scale. On the other hand, the *slow scale* is related to the time span over which anomalies may occur and exhibit non-stationary evolution of the system dynamics. SDF detects the statistical changes in behavioral patterns over slow-scale epochs that are simply referred to as epochs in the sequel. The first part of the method, which is called the *forward problem* involves generation of patterns from training data, which is comprised of the following steps.

- Sensor time series data, generated from a physical system or its dynamical model, are collected at several training epochs over the range of operation. A compact (i.e., closed and bounded) region $\Omega \in \mathbb{R}^n$, where $n \in \mathbb{N}$, within which the stationary time series is circumscribed, is identified. Let the space of time series data sets be represented as $\mathcal{Q} \subseteq \mathbb{R}^{n \times N}$, where $N \in \mathbb{N}$ is sufficiently large for convergence of statistical properties within a specified threshold. Then, $\{\mathbf{q}^i\} \in \mathcal{Q}$ denotes a time series at an epoch $i \in \{0, 1, ..., l-1\}$, where l is the number of epochs under consideration. Let the epoch 0 denote the reference/nominal condition.
- Encoding of Ω is accomplished by introducing a partition $\mathbb{B} \equiv \{B_0, ..., B_{(m-1)}\}$ consisting of m mutually exclusive (i.e., $B_j \cap B_k = \emptyset \ \forall j \neq k$), and exhaustive (i.e., $\cup_{j=0}^{m-1} B_j = \Omega$) cells. Let, each cell be labeled by symbols $s_j \in \Sigma$ where $\Sigma = \{s_0, ..., s_{m-1}\}$ is called the alphabet. This process of coarse graining can be executed by uniform, maximum entropy, or any other form of partitioning with respect to the data set at epoch 0. Then, the time series data points $\{\mathbf{q}^0\}$ that visit the cell B_j are denoted as $s_j \ \forall j = 0, 1, ..., m-1$. This step enables transformation of the reference time series data $\{\mathbf{q}^0\}$ to a symbol sequence $\{\mathbf{s}^0\}$. To alleviate the difficulties in partitioning of noisy time series, it can be transformed to wavelet space [8] or analytic signal space [7]. Similar to the reference condition, symbol sequences $\{\mathbf{s}^0\}, \{\mathbf{s}^1\}, ..., \{s^{l-1}\}$ are generated from the data at other epochs using the same partitioning created at the reference condition.
- A probabilistic finite state machine (PFSA) is then constructed with a chosen depth and the training epoch symbol sequences are run through it. Thus a state transition matrix $\Pi^i = [\pi^i_{jk}]$, where $j, k \in \{1, 2, ..., r\}$ are the states of the PFSA with an $(r \times r)$ state transition matrix, is obtained for each training epoch i. Since $\pi^i_{jk} \geq 0$ is the transition probability from state j to state k, Π^i is a stochastic matrix, i.e., $\sum_k \pi^i_{jk} = 1$. Often to compress the information further, the state probability vector $\mathbf{p}^i = [p^i_1 \ \cdots \ p^i_r]$ that is the left eigenvector corresponding to the (unique) unity eigenvalue of the irreducible stochastic matrix Π^i) is calculated at each epoch i. The vector $\mathbf{p}^i$ is called the pattern vector at the respective training epoch i in the sequel, and is a relatively low-dimensional representation of the dynamical system at the current epoch. Let $\mathcal{P}$ denote the space of pattern vectors $\mathbf{p}$.
- The anomaly measure, also called deviation measure, is obtained from the distance of a pattern vector $\mathbf{p}^i$ (or state transition matrix Π^i) at an epoch i from the reference pattern vector $\mathbf{p}^0$ (or state transition matrix Π^0) at epoch 0. For example, the deviation measure μ at epoch i can be calculated as $\mu(\mathbf{p}^i, \mathbf{p}^0)$ as a scalar distance (e.g. Euclidean norm or Kullback distance [11]) between two vectors $\mathbf{p}^i$ and $\mathbf{p}^0$. In the forward problem, evolution of deviation patterns over the epochs is determined in a statistical sense [12].

Given a deviation measure, the *inverse problem* yields an estimate of the current condition of the system.

An example of anomaly detection for a nonlinear electronic system apparatus modeled on the Duffing equation [5], [13] is considered at this point. The Duffing equation is given below.

$$\frac{d^2y}{dt^2} + \beta\frac{dy}{dt} + y(t) + y^3(t) = Acos(\Omega t) \tag{1}$$

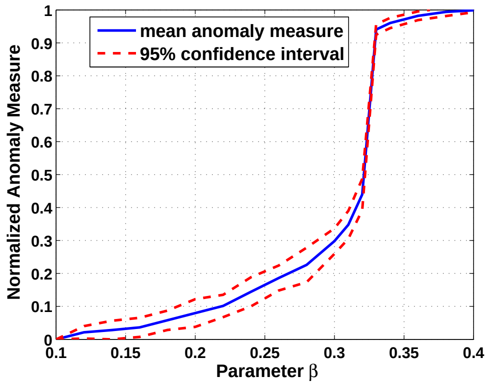

Fig. 1. Anomaly Evolution Profile for Duffing equation

where the dissipation parameter β varies slowly with respect to the response of the dynamical system; and a change in the value of β is considered as an anomaly. With the input amplitude $A = 22.0$ and input frequency $\Omega = 5.0$, Fig. 1 shows the anomaly evolution profile obtained from the forward problem using maximum entropy partitioning with alphabet size 8.From the perspective of inverse problem, monotonicity of anomaly evolution is a necessary condition for identification of β based on a given value of the deviation measure. However, in the incipient anomaly condition the slope of the curve is observed to be small and due to noise it can be seen that the anomaly evolution curve has significant variance, which causes a potential difficulty in the execution of the inverse problem, i.e., given a deviation measure, there may exist a wide range of possible values of β.

The SDF-based anomaly detection tool has been shown to be superior in performance, especially for detection of incipient anomalies, compared to other pattern recognition tools, such as principal component analysis (PCA), artificial neural network (ANN), and particle filter (PF). In this context, the focus of this paper is to explore further improvement of the SDF performance for the purpose of anomaly detection. In the current SDF methodology, and also in other literature, partitioning is done based on nominal data. In such cases, even if the partitioning is optimal (e.g., in terms of maximum entropy or some other criteria) under nominal conditions, there is no guarantee that it will satisfy the monotonicity condition. On the other hand, partitioning based on the nominal stationary data does not consider evolutionary characteristics of the data over the epochs. Hence, it may be advantageous to take non-stationary dynamics into consideration and create partitioning based on changes of time series data over different epochs. This is the key idea of the maximum migration partitioning (MMP) method; the rationale of this nomenclature is explained in the next section. The ultimate goal is to obtain a well-conditioned anomaly evolution curve even at the incipient anomaly conditions, which is very useful for anomaly (or fault) detection in real complex systems.

A. The Maximum Migration Partitioning (MMP)

Given the structure of the PFSA, a mapping from the data space to the pattern space, denoted by $\Gamma_{\mathbb{B}}(\cdot)$, depends only on the chosen partitioning $\mathbb{B}$. Let pattern vectors $\mathbf{p}^i$ be considered for anomaly detection. Hence, $\Gamma_{\mathbb{B}}(\mathbf{q}^i) = \mathbf{p}^i_{\mathbb{B}}$ for epoch $i \in \{0, 1, ..., l-1\}$, where $\mathbf{q}^i$ is the time series data as defined earlier in Section 2. Furthermore, let the epoch be a function of a single system parameter. Then, the maximum migration partitioning (MMP) $\mathbb{B}^*$ satisfies the following two conditions.

1) The mapping $\Gamma_{\mathbb{B}^*} : \mathcal{Q} \rightarrow \mathcal{P}$ is bijective, i.e. time series data at two different epochs that are statistically different are not be mapped as the same pattern; otherwise, it will be impossible to distinguish between deviation measures at those two epochs in the inverse problem. Therefore, bijectivity is a necessary condition for MMP. Since the deviation measure μ is a continuous function of the pertinent slowly varying parameter(s), bijectivity of $\Gamma_{\mathbb{B}^*}$ implies strict monotonicity and is a necessary condition for MMP.
2) With the constraint of the first condition, the $\mathbb{B}^*$ is given by maximizing the multi-objective vector reward function $J_{\mathbb{B}}$ defined as

$$\mathbb{B}^* = \arg\max_{\mathbb{B}} J_{\mathbb{B}} \triangleq \arg\max_{\mathbb{B}} \begin{bmatrix} \mu(\mathbf{p}^1_{\mathbb{B}}, \mathbf{p}^0_{\mathbb{B}}) \\ \mu(\mathbf{p}^2_{\mathbb{B}}, \mathbf{p}^0_{\mathbb{B}}) \\ \vdots \\ \mu(\mathbf{p}^{l-1}_{\mathbb{B}}, \mathbf{p}^0_{\mathbb{B}}) \end{bmatrix} \quad (2)$$

Condition 2 justifies the nomenclature of the maximum migration partitioning (MMP) technique as explained in the following remark.

Remark 2.1: At an off-nominal epoch i, i.e., $i > 0$, net migration of data points from one cell of the partitioning to another is maximized when the time series evolves from $\mathbf{q}^0$ to $\mathbf{q}^i$. That is why this partitioning is named maximum migration. The concept of MMP is analogous to the that of particle migration from one energy level to another in the Statistical Mechanics setting [14]; however, only the net data point migration is relevant to partitioning. Otherwise, migration of data points among different partitioning cells, which keeps the state occupation probabilities unchanged, will result in a zero deviation measure. Therefore, in the context of Statistical Mechanics, the particles are identical and indistinguishable. If evolution of the state transition matrix Π is considered instead of the pattern vectors $\mathbf{p}$, it might be possible to identify a change in Π without any net migration of data points.

The satisfaction of the above two conditions may require Pareto optimization [15] even for one-dimensional time series data sets. It is possible for certain ill-conditioned data sets the solution of the optimization procedure may become intractable. While the future work will be devoted to develop an elaborate optimization algorithm, this paper provides a relatively simpler algorithm for one-dimensional (i.e., $n = 1$) time-series data in the following subsection, primarily to establish validity of the proposed MMP concept.

B. Maximum Migration Partitioning Procedure

This section describes a sub-optimal solution for the constrained multi-objective optimization problem described in Section 2-A. The objective space $\mathcal{O}$ consists of the reward vector J, while decisions are made in the space $\mathcal{P}$ of all possible partitions. The reward vector J is broken up into individual rewards functions $J^i = \mu(\mathbf{p}^i, \mathbf{p}^0)$, that are independently optimized to obtain the individual maxima. As a further simplification, it is assumed that the time series data is one-dimensional (i.e., $n = 1$) wherein a partition consisting of m cells may be succinctly represented by the $m-1$ points that separate the cells. In other words, an m-cell partition is expressed by $\Lambda \triangleq \{\lambda_1, \lambda_2, \cdots, \lambda_{m-1}\}$ with cardinality $|\Lambda| = m - 1$.

The deviation measure for the i^{th} epoch $\mu(\mathbf{p}^i, \mathbf{p}^0)$ is dependent on its specific partition Λ and is denoted by $J^i(\Lambda) = \mu(\mathbf{p}^i_\Lambda, \mathbf{p}^0_\Lambda)$. This sub-optimal partitioning scheme involves sequential estimation of the elements of the partitioning Λ.

The partitioning process is initiated by dividing the data set into two cells, $\Lambda_2 = \{\lambda_1\}$, where λ_1 is evaluated as

$$\lambda_1^* = \arg\max_{\lambda_1} J^i(\Lambda_2) = \arg\max_{\lambda_1} \mu(\mathbf{p}^i_{\Lambda_2}, \mathbf{p}^0_{\Lambda_2}) \qquad (3)$$

Now, the two-cell optimal partitioning is given by $\Lambda_2^* = \{\lambda_1^*\}$. The next step is to partition the data into three cells as Λ_3) by dividing either of the two existing cells of Λ_2^* with the placement of a new partition boundary at λ_2, where λ_2 is evaluated as

$$\lambda_2^* = \arg\max_{\lambda_2} J^i(\Lambda_3) = \arg\max_{\lambda_2} \mu(\mathbf{p}^i_{\Lambda_3}, \mathbf{p}^0_{\Lambda_3}) \qquad (4)$$

where $\Lambda_3 = \{\lambda_1^*, \lambda_2\}$. The optimal 3-cell partitioning is obtained as $\Lambda_3^* = \{\lambda_1^*, \lambda_2^*\}$. In this (local) optimization procedure, the cell that provides the largest increment in the reward function upon further segmentation ends up being partitioned. Iteratively, this procedure can be extended to obtain the m cell partition as follows.

$$\lambda_{m-1}^* = \arg\max_{\lambda_{m-1}} J^i(\Lambda_m) = \arg\max_{\lambda_{m-1}} \mu(\mathbf{p}^i_{\Lambda_m}, \mathbf{p}^0_{\Lambda_m}) \qquad (5)$$

where $\Lambda_m = \Lambda_{m-1}^* \cup \{\lambda_{m-1}\}$ and the optimal m cell partitioning is given by $\Lambda_m^* = \Lambda_{m-1}^* \cup \{\lambda_{m-1}^*\}$

This optimization procedure is monotonically increase in the reward function with every additional sequential operation, i.e. $J^i(\Lambda_{m-1}^*) \leq J^i(\Lambda_m^*)$. This is evident from the following argument.

Let Λ_{m-1}^* be the $(m-1)$-cell partition that maximizes the reward $J^i(\Lambda_{m-1})$. Based on the algorithm $\Lambda_m = \Lambda_{m-1}^* \cup \{\lambda_{m-1}\}$), if λ_{m-1} is chosen such that it already belongs to Λ_{m-1}^*, then there would be no change in the partitioning structure and $J^i(\Lambda_m) = J^i(\Lambda_{m-1}^*)$. Since $J^i(\Lambda_m^*) \geq J^i(\Lambda_m)\ \forall \Lambda_m$, it follows that $\max J^i(\Lambda_{m-1}) \leq \max J^i(\Lambda_m)$. The monotonicity in the reward function J^i allows formulation of a rule for termination of the sequential optimization algorithm. The process of creating additional partitioning cells is stopped if the gain in the reward falls below a specified threshold η_{stop} and the stopping rule is:

$$J^i(\Lambda_m^*) - J^i(\Lambda_{m-1}^*) \leq \eta_{stop}. \qquad (6)$$

Let the suboptimal maximum migration partitioning (MMP) at an epoch i be denoted as $\Lambda^*(i)$, in which the above procedure is followed to obtain consecutive partitions $\Lambda^*(1), \Lambda^*(2), ..., \Lambda^*(l-1)$ for $l-1$ off-nominal training epochs. These partitions capture the slowly varying non-stationary evolution of the dynamical system at different epochs. If the statistical characteristics of the dynamical system do not significantly change over the range of the training epochs, then it is expected that $\Lambda^*(i) \approx \Lambda^*(j)$ for $i \neq j$. In other words, the cell boundaries in the partitions for different epochs form clusters on the decision space, which is viewed as a near-Utopian global partitioning Λ^*_{global} that takes the mean of partitions across all training epochs to simultaneously satisfy optimization criteria for different epochs; it is considered as the universal partitioning for the data set (e.g., the data sets of the Duffing system in Section 3).

Remark 2.2: However, there may exist certain epochs for which the partitioning is significantly different from the rest. In this paper, these outlier partitions are not considered for construction of Λ^*_{global}. In general, if the partitioning for individual epochs are not clustered in the decision space, a Pareto surface should be constructed. Furthermore, since this method of partitioning may not guarantee monotonicity. it needs to be chosen from the Pareto set based on the monotonicity constraint (see Section 2-A) and weights assigned to each epoch. This is a topic of future research.

3. Experimental Validation of Maximum Migration Partition

This section evaluates the performance of maximum migration partitioning (MMP) for anomaly detection by experimentation on a nonlinear electronic circuit that implements Duffing equation.

A. Anomaly detection in Duffing system

Experiments have been conducted for anomaly detection in a nonlinear active electronic circuit that emulates the forced Duffing equation [10]

$$\frac{d^2y}{dt^2} + \beta\frac{dy}{dt} + y(t) + y^3(t) = Acos(\Omega t) \qquad (7)$$

where amplitude $A = 22.0$, $\Omega = 5.0$ and the dissipation parameter β is slowly varied from 0.1 to 0.4. The nominal condition is at $\beta = 0.1$ and the anomalous conditions are at $\beta = 0.20, 0.26$ and 0.32. A detailed description of the experimental setup and the accompanying results for anomaly detection obtained by using the maximum entropy partitioning (MEP) scheme reported in earlier publications [8][12]. This subsection compares the results obtained by MEP with those obtained by MMP that is developed in this paper.

The time series $\mathbf{q}^0, \mathbf{q}^1, \mathbf{q}^2$ and $\mathbf{q}^3$, sampled from the sensor data $y(t)$ in the experiment, are collected for the nominal condition at$\beta = 0.10$ and for three other training epochs representing anomalous conditions at $\beta = 0.20$, $\beta = 0.26$ and $\beta = 0.32$, respectively. The MMP scheme is then implemented to obtain the partition sets $\Lambda^*(1), \Lambda^*(2)$ and $\Lambda^*(3)$ for the three anomalous epochs, respectively.

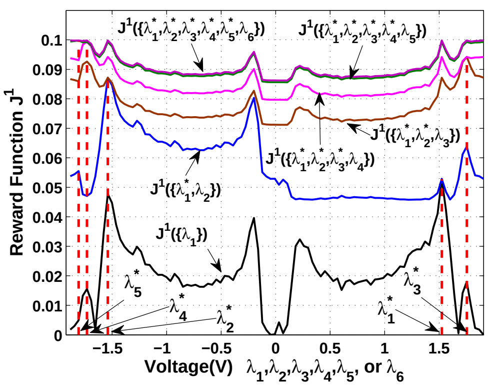

Fig. 2. Reward function vs. partitions for Epoch 1 in Duffing system

Figure 2 depicts the optimization process for obtaining the partition set $\Lambda^*(1)$ at $\beta = 0.20$, where λ_1^* is evaluated by maximizing the reward function $J^1(\{\lambda_1\})$. Figure 2 shows both the reward function and its corresponding optimal value λ_1^* and the second partition λ_2^* is obtained by maximizing the function $J^1(\{\lambda_1^*, \lambda_2\})$. As described in Subsection 2-A, λ_1^* is kept fixed while λ_2 is optimized. This suboptimal process is recursively continued until the threshold $\eta_{stop} = 4.0 \times 10^{-4}$ is reached, which leads to the creation of 6 cells (i.e., 5 partitions) denoted by $\Lambda_6^*(1) = \{\lambda_1^*, \cdots, \lambda_5^*\}$ as shown in Fig. 2.

The above procedure is repeated for the two other anomalous epochs at $\beta = 0.26$ and $\beta = 0.32$ to obtain the partition set $\Lambda_6^*(2)$ and $\Lambda_6^*(3)$, respectively. These three partition sets are superimposed on the time series data in Fig. 3. Interestingly, even though the three partition sets were obtained by independent maximization of the reward functions $J^1 = \mu(\mathbf{p}^1, \mathbf{p}^0)$, $J^2 = \mu(\mathbf{p}^2, \mathbf{p}^0)$ and $J^3 = \mu(\mathbf{p}^3, \mathbf{p}^0)$; they lie in close proximity of one another. In fact the partitions obtained across the epochs are within a bound of $\epsilon = 0.15$ from each other. This result justifies the conjecture made at the end of subsection 2-B and eludes to the possibility of obtaining a near-utopian (sub)optimal partitioning scheme. The mean of the three partition set is taken as the final partitioning Λ_{global}^*, obtained by the MMP method. The alphabet size is the same as the number of partitioning cell and is equal to 6.

For maximum entropy partitioning (MEP) [8], an alphabet size of $|\Sigma| = 8$ is chosen and the partition set is generated

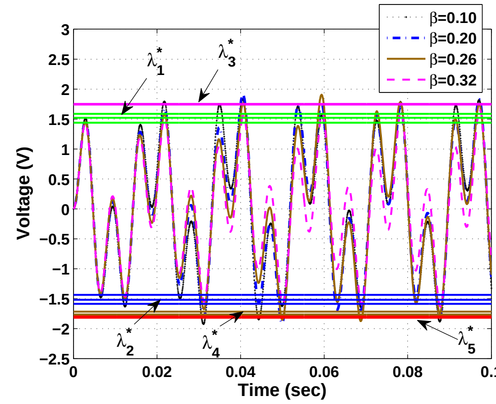

Fig. 3. Partitioning Duffing time series data using MMP

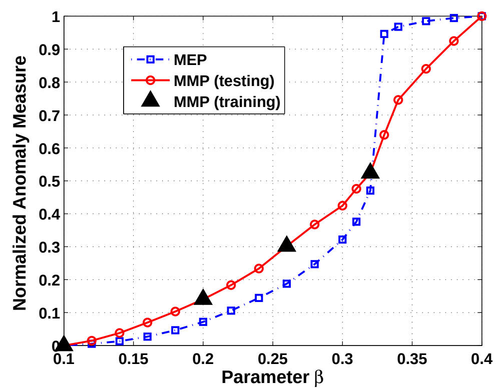

Fig. 4. Comparison of anomaly evolution profiles

from the nominal data $\beta = 0.10$. Normalized measure $\mu(\mathbf{p}^i, \mathbf{p}^0)$ quantifies the deviation of the Duffing system behavior from its nominal condition. Euclidean distance is used as $\mu(\cdot, \cdot)$ for the whole exercise. Figure 4 plots the deviation measure versus the dissipation parameter β by using both the MEP and the MMP methods. The following qualitative differences between the deviation measure profiles obtained by the two different partitioning scheme are observed.

- The deviation measure μ in MMP provides high sensitivity in the region close to $\beta = 0.1$; hence, a better detection of incipient faults becomes possible along with reduction in probability of false alarms and missed detections of anomaly. It also provides a better estimation of the parameter β.
- The Duffing system is nonlinear and undergoes bifurcations. Phase plots and time response plots in [13] show that the system behavior abruptly changes around $\beta = 0.32$. The partitioning obtained by MEP depends only on the nominal condition ($\beta = 0.1$) and is therefore rendered ineffective for anomaly detection for β is beyond the point of bifurcation. In fact, Fig. 4 shows that the sensitivity of the MEP anomaly profile is almost negligible for large values of β (i.e., > 0.32). On the other hand, the MMP deviation measure profile has

sufficient sensitivity in the region beyond the point of bifurcation. This may be attributed to the fact that MMP captures the evolution of the dynamical system from a training set and, unlike MEP, it is not based on any single epoch.

Based on the above observations, it is concluded that MMP has certain advantages over MEP for partitioning, especially for detection of incipient anomaly and for estimation of critical parameter(s) beyond the point of bifurcation.

4. Summary, Conclusions and Future work

This paper presents a novel method of partitioning time series data for anomaly detection. In this approach, the transformation from the data space to pattern space is optimized to improve the capability of symbolic feature extraction. The advantage of using maximum migration partitioning (MMP) over maximum entropy partitioning (MEP) are demonstrated by testing on experimental data. The major aspects of this methodology is summarized below along with important conclusions.

- *Identification of the evolving characteristics of the dynamical system*: The MMP is shown to capture the nonstationary evolution of a dynamical system. Hence, it has the capability to generate a well-conditioned anomaly evolution profile in the forward problem that leads to better anomaly estimation in the inverse problem.
- *Optimization of alphabet size*: During traditional partitioning process, the alphabet size is typically a user-defined quantity. The stopping rule (see Subsection 2-A) in the MMP scheme provides a way to systematically arrive at the correct alphabet size. This helps in avoiding redundant partitions, which thereby reduces computational complexity.
- *Issue of Sensitivity and Robustness*: In the partitioning algorithm, MMP is made sensitive to small changes in the time series. Although this property improves the anomaly detection capability for incipient anomalous conditions, robustness may become a potential problem, especially for data with high noise content. As discussed in Subsection 2-B and Section 3, over different epochs, partitions are observed to form clusters of certain widths. It is also observed that fluctuation of partitions within that cluster does not change the anomaly evolution profile significantly, proving the method's robustness property. However, further in depth study on the tradeoff between sensitivity and robustness will be an important future work.
- MMP is shown to be a tool to magnify the deviations in data space during transformation to pattern space and this is achieved by rewarding larger distances between pattern vectors in pattern space. However, another potential promise of MMP is to make the pattern space input independent, which will make nonlinear systems behave like a linear system. This might be possible by penalizing the deviations in pattern space upon change in input for a particular epoch.

As the concept of MMP is proposed here for the first time to the best of the authors' knowledge, this method of space partitioning for anomaly detection requires continued theoretical and experimental research. The following topics are recommended for future research.

- Formulation of a Pareto multi-objective optimization algorithm to generate MMP for a general time-series data.
- Extension of the MMP algorithm for multi-dimensional time-series data.
- Modification of the MMP algorithm in transformed space (e.g., Hilbert transform, wavelet transform) of time-series data.
- Modification of the MMP algorithm for input independence by penalizing the deviations in pattern space upon change in input for a particular epoch.

References

[1] K. Fukunaga, *Statistical Pattern Recognition, 2nd Edition*. Academic Press, Boston, USA, 1990.

[2] T. Lee, *Independent component analysis: Theory and applications*. Kluwer Academic Publishers, Boston, USA, 1998.

[3] R. Rosipal, M. Girolami, and L. Trejo, "Kernel pca feature extraction of event-related potentials for human signal detection performance," *Proc. Int. Conf. Artificial Neural Networks Medicine Biol.*, pp. 321–326, 2000.

[4] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-04), Washington D.C.*, 2004.

[5] A. Ray, "Symbolic dynamic analysis of complex systems for anomaly detection," *Signal Processing*, vol. 84, no. 7, pp. 1115–1130, July 2004.

[6] R. Steuer, L. Molgedey, W. Ebeling, and M. Jimenez-Montano, "Entropy and optimal partition for data analysis," *The European Physical Journal B*, vol. 19, pp. 265–269, 2001.

[7] A. Subbu and A. Ray, "Space partitioning via Hilbert transform for symbolic time series analysis," *Applied Physics Letters*, vol. 92, no. 8, pp. 084107–1 to 084107–3, 2008.

[8] V. Rajagopalan and A. Ray, "Symbolic time series analysis via wavelet-based partitioning," *Signal Processing*, vol. 86, no. 11, pp. 3309–3320, Nov 2006.

[9] S. Sarkar, K. Mukherjee, and A. Ray, "Generalization of Hilbert transform for symbolic analysis of noisy signals," *Signal Processing*, pp. in press, DOI: 10.1016/j.sigpro.2008.12.009, 2008.

[10] J. Thompson and H. Stewart, *Nonlinear Dynamics and Chaos*. Wiley, Chichester, UK, 1986.

[11] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[12] S. Gupta and A. Ray, "Real-time fatigue life estimation in mechanical structures," *Measurement Science and Technology*, vol. 18, pp. 1947–1957, May 2007.

[13] C. Rao, A. Ray, S. Sarkar, and M. Yasar, "Review and comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns," *Signal Processing*, p. in press, 2008.

[14] R. Pathria, *Statistical Mechanics, 2nd ed.* Mutterworth Heinmann, Oxford, UK, 1996.

[15] R. Steuer, *Multiple Criteria Optimization: Theory, Computations, and Application*. John Wiley & Sons, Inc., New York, USA, 1986.