SIViP (2016) 10:1535–1542
DOI 10.1007/s11760-016-0967-5

ORIGINAL PAPER

Symbolization of dynamic data-driven systems for signal representation

Soumalya Sarkar[1,2] · **Pritthi Chattopdhyay**[1] · **Asok Ray**[1]

Received: 29 February 2016 / Revised: 20 July 2016 / Accepted: 9 August 2016 / Published online: 19 August 2016

Abstract The underlying theory of symbolic time series analysis (STSA) has led to the development of signal representation tools in the paradigm of dynamic data-driven application systems (DDDAS), where time series of sensor signals are partitioned to obtain symbol strings that, in turn, lead to the construction of probabilistic finite state automata (PFSA). Although various methods for construction of PFSA from symbol strings have been reported in literature, similar efforts have not been expended on identification of an appropriate alphabet size for partitioning of time series, so that the symbol strings can be optimally or suboptimally generated in a specified sense. The paper addresses this critical issue and proposes an information-theoretic procedure for partitioning of time series to extract low-dimensional features, where the key idea is suboptimal identification of boundary locations of the partitioning segments via maximization of the mutual information between the state probability vector of PFSA and the members of the pattern classes. Robustness of the symbolization process has also been addressed. The proposed alphabet size selection and time series partitioning algorithm have been validated by two examples. The first example addresses parameter identification in a simulated Duffing system with sinusoidal input excitation. The second example is built upon an ensemble of time series of chemiluminescence data to predict lean blowout (LBO) phenomena in a laboratory-scale swirl-stabilized combustor apparatus.

Keywords Symbolic time series analysis · Information theory · Probabilistic finite state automata

✉ Asok Ray
axr2@psu.edu

Soumalya Sarkar
sms388@gmail.com

Pritthi Chattopdhyay
pritthichatterjee@gmail.com

[1] Department of Mechanical and Nuclear Engineering, The Pennsylvania State University, University Park, PA 16802, USA

[2] Present Address: United Technology Research Center, East Hartford, CT, USA

1 Introduction

Symbolic time series analysis (STSA) [1–3] has been used for constructing statistical models of (possibly nonlinear) dynamical systems, which rely on temporal and spatial discretization based on the fundamental concepts of symbolic dynamics [4]. Along this line, Ray and coworkers [5–8] have developed data-driven procedures for generation of probabilistic finite state automata (PFSA) models, where the major role of STSA is to serve as a feature extraction tool for information compression and pattern classification in dynamic data-driven application systems (DDDAS) (see [9] and references therein for details of the DDDAS concept); this procedure has been used for early detection of anomalous behavior as well as for pattern recognition in diverse physical systems (e.g., see [10,11]). While extensive research work (e.g., see [12]) has been reported for investigation of the properties and variations of transformation from a symbol space to a feature space in the conversion of symbol strings into PFSA models, similar efforts have not been expended on how to find an optimal alphabet size for symbolization of time series (e.g., see [13–15]).

The paper develops an information-theoretic procedure of time series partitioning in the paradigm of dynamic data-driven systems, where the objective is to extract low-dimensional features from time series for pattern classification via construction of probabilistic finite state automata

(PFSA). In particular, symbol sequences are constructed from sensor time series of the underlying process in such a way that the (relatively slowly evolving) statistical changes are captured over a given set of training data belonging to different classes. The key idea lies in optimal partitioning of the time series via maximization of the mutual information [16] between the state probability vector (treated as a feature [5]) of a PFSA and the members of the pattern classes.

Major contributions of this paper are stated below.

1. Partitioning of time series in a way that maximizes the mutual information [16] between the symbolic dynamic feature (e.g., the state probability vector of PFSA) and the pattern class to which it belongs.
2. Development of robust algorithms of alphabet size selection for symbolization of time series so that the symbol strings can be optimally generated in a specified sense.
3. Performance comparison of the proposed partitioning technique with another commonly used method, namely maximum entropy partitioning (MEP).
4. Validation of the proposed concepts on simulated data from a sinusoidally excited Duffing system, and experimental data from a laboratory-scale swirl-stabilized combustor for lean blowout (LBO) prediction.

2 Motivation and preliminaries of STSA

This section briefly presents the preliminaries and concepts of symbolic time series analysis (STSA) [2,5,8] and related information-theoretic definitions [16].

Partitioning for Symbolization: Stauer et al. [17] reported a comparison of maximum entropy partitioning and uniform partitioning, where it is concluded that maximum entropy partitioning is, in general, a better tool for change detection in symbolized time series than uniform partitioning. Buhl and Kennel [13] reported symbolic false nearest neighbor partitioning (SFNNP) to optimize generating partitions by avoiding topological degeneracy. However, SFNNP suffers from high computational complexity and low robustness to noise. Rajagopalan and Ray [6] introduced wavelet space partitioning (WSP), where the wavelet transform largely alleviates the above shortcoming and is particularly effective with noisy data. Subbu and Ray [7] introduced Hilbert-transform-based analytic signal space partitioning (ASSP) as an alternative to WSP. Nevertheless, these techniques emphasizes on modeling more than anomaly detection. Jin et al. [18] reported the theory and validation of a wavelet-based feature extraction tool that used maximum entropy partitioning of the space of wavelet coefficients. Even if this partitioning is optimal (e.g., in terms of maximum entropy or some other criteria) under nominal conditions, it may not remain optimal at other conditions. Along this line Sarkar et al. [14] proposed a time series partitioning procedure to extract low-dimensional features from time series while optimizing the class separability; however, this method is strongly dependent on the choice of the classifier tool.

The goal of the work, reported in the current paper, is to overcome the difficulties of the above-mentioned partitioning methods with the objective of making STSA a robust data-driven feature extraction tool based on an information-theoretic concept. Symbol string generation from time series data is posed as a two-time-scale problem. The *fast scale* is related to the response time of the process dynamics. In contrast, the *slow scale* is related to the time span over which non-stationary evolution of the system dynamics may occur. Encoding of the data space of time series is accomplished by introducing a partition that consists of finitely many mutually exclusive and exhaustive cells. Let the jth cell be labeled by a symbol $\sigma_j \in \Sigma$ and data points of the time series, which visit the jth cell, are denoted as σ_j. The finite set $\Sigma = \{\sigma_0, ..., \sigma_{|\Sigma|-1}\}$ is called an alphabet, and its cardinality $|\Sigma| \geq 2$ is called the alphabet size.

Construction of PFSA: A probabilistic finite state automaton (PFSA) is constructed from the symbol string and is modeled as a quadruple $K = (\Sigma, Q, \delta, \pi)$, where

- The symbol alphabet Σ is a (nonempty) finite set;
- The set Q of automaton states is (nonempty) finite;
- The state transition function $\delta : Q \times \Sigma \to Q$;
- The morph function $\pi : Q \times \Sigma \to [0, 1]$, where $\sum_{\sigma \in \Sigma} \pi(q, \sigma) = 1$ for all $q \in Q$. The morph function π generates the $\left(|Q| \times |\Sigma|\right)$ morph matrix Π. To compress the information further, the state probability vector of the PFSA is used as an extracted feature.

Construction of D-Markov Machines: A D-Markov machine [5,8] (a type of PFSA) is a statistically stationary symbol string $\ldots s_{-1}s_0s_1 \ldots$, where each s_i is a symbol in Σ and the probability of occurrence of a new symbol depends only on the last D symbols (Depth), i.e.,

$$P[s_n \mid \ldots s_{n-D} \ldots s_{n-1}] = P[s_n \mid s_{n-D} \ldots s_{n-1}] \tag{1}$$

It is noted that D-Markov machines belong to the class of shifts of finite type [4]. In a D-Markov machine, a word $w \in \Sigma^D$ can be associated with a state of the machine. Let ν_{ij} be the number of times that a symbol σ_j is generated from the state q_i upon observing a symbol string. The maximum a posteriori probability (MAP) estimate of the probability map for the PFSA is computed by frequency counting [8] as:

$$\hat{\pi}_{MAP}(q_i, \sigma_j) \triangleq \frac{1 + \nu_{ij}}{|\Sigma| + \sum_{\ell=1}^{|\Sigma|} \nu_{i\ell}} \tag{2}$$

The rationale for initializing each element of the count matrix to 1 is that: If no event is generated at a state $q \in Q$, then there should be no preference to any particular symbol. Then, $\hat{\pi}_{MAP}(q, \sigma) = \frac{1}{|\Sigma|}\ \forall \sigma \in \Sigma$, i.e., uniform distribution of event generation at the state q.

3 Problem formulation

The success of a time series partitioning methodology depends on how much information embedded in the time series is captured by the PFSA (e.g., the state probability vector $\boldsymbol{p}$). A lower bound of the probability P_E of incorrect estimation $\hat{\mathscr{C}}$ of the true class $\mathscr{C}$ (i.e., $P_E = Pr[\hat{\mathscr{C}} \neq \mathscr{C}]$) is obtained from the weak form of Fano's inequality [16] as:

$$P_E \geq \frac{H(\mathscr{C}|\mathscr{P}) - 1}{\log_2 |\mathscr{C}|} = \frac{H(\mathscr{C}) - I(\mathscr{P}; \mathscr{C}) - 1}{\log_2 |\mathscr{C}|} \tag{3}$$

where the random vector $\mathscr{P}$ represents the input feature extracted from a given time series whose pattern class (belonging to $\mathscr{C} = \{c_0, c_1, \ldots, c_{(|\mathscr{C}|-1)}\}$) may be unknown; $H(\mathscr{C})$ and $H(\mathscr{C}|\mathscr{P})$ are the entropy and conditional entropy of pattern class $\mathscr{C}$, respectively; and $I(\mathscr{P}; \mathscr{C})$ is the mutual information between the input feature and the pattern class. Since $H(\mathscr{C})$ and $|\mathscr{C}|$ are fixed, the lower bound of P_E is minimized when $I(\mathscr{P}; \mathscr{C})$ reaches its maximum.

As the continuity and higher order differentiability of the partitioning function in the range space of mutual information is neither guaranteed nor adequately analyzed, a sequential search-based technique has been adopted for optimized alphabet size selection instead of a gradient-based optimization procedure. The process of constructing a search space is started with an initial fine grid size, where each of the grid boundaries denotes a possible cell boundary of trial partitions. These boundaries can be obtained, e.g., via either uniform partitioning (UP) or maximum entropy partitioning (MEP) [6] of the range space of the time series.

At the beginning, the time series is divided into L regions on its signal space that is marked by $(L-1)$ boundaries Υ (excluding the end points) for search via MEP or UP. The positive integer L, where $L > |\Sigma|$, is to be selected by the user. Thus, for an alphabet Σ, there are $(|\Sigma| - 1)$ partitioning boundaries to choose from $(L-1)$ possibilities, which is equivalent to making one choice in the space of all possible partitioning vectors (i.e. selecting one $(|\Sigma| - 1)$-dimensional partitioning vector from ${}^{(L-1)}C_{(|\Sigma|-1)}$ different choices). It follows that the space $\mathcal{P}$ of all possible partitioning boundaries may become significantly large as L and $|\Sigma|$ are increased. For example, if $L \gg |\Sigma|$, then computational complexity increases approximately by a factor of $L/|\Sigma|$ as $|\Sigma|$ is increased by one. The cost function to be maximized is the scalar-valued mutual information $I(\mathscr{P}; \mathscr{C})$, while decisions are made in the space $\mathcal{P}$ of partitioning boundaries. The cost is dependent on a specific partitioning Λ, because the extracted feature $\mathscr{P}$ is a function of the chosen $(|\Sigma| - 1)$-dimensional partitioning vector Λ; the cost is denoted by $I(\mathscr{P}(\Lambda); \mathscr{C})$. This suboptimal partitioning scheme involves sequential estimation of the elements of the partitioning vector Λ.

The partitioning process is initiated by searching the optimal cell boundary that divides the data range into two cells, i.e., $\Lambda_2 = \{\lambda_1\}$, where λ_1 is optimized as:

$$\lambda_1^* = \arg\max_{\lambda_1 \in \Upsilon} I(\mathscr{P}(\Lambda_2); \mathscr{C}) \tag{4}$$

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

$$\lambda_2^* = \arg\max_{\lambda_2 \in \Upsilon \setminus \Lambda_2^*} I(\mathscr{P}(\Lambda_3); \mathscr{C}) \tag{5}$$

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 $I(\mathscr{P}; \mathscr{C})$ upon further segmentation ends up being partitioned. Iteratively, this procedure is extended to obtain the parameter $|\Sigma|$ of cell partitioning as follows:

$$\lambda_{|\Sigma|-1}^* = \arg\max_{\lambda_{|\Sigma|-1} \in \Upsilon \setminus \Lambda_{|\Sigma|-1}^*} I(\mathscr{P}(\Lambda_{|\Sigma|}); \mathscr{C}) \tag{6}$$

where $\Lambda_{|\Sigma|} = \Lambda_{|\Sigma|-1}^* \cup \{\lambda_{|\Sigma|-1}\}$ and the optimal $|\Sigma|$ is given by $\Lambda_{|\Sigma|}^* = \Lambda_{|\Sigma|-1}^* \cup \{\lambda_{|\Sigma|-1}^*\}$.

In this optimization procedure, the mutual information increases monotonically with additional sequential operation provided that the mutual information is computed correctly. This monotonicity property allows formulation of a rule for termination of the sequential optimization algorithm. The process of creating additional partitioning cells is stopped if the normalized mutual information, relative to $H(\mathscr{C})$ with a uniform class prior, crosses a specified positive threshold $I_{max} \in [0, 1]$. The stopping criterion is: $\Lambda_{|\Sigma|}^*$ is the optimal partitioning and $|\Sigma|$ is the optimal alphabet size if

$$\frac{I(\mathscr{P}(\Lambda_{|\Sigma|}^*); \mathscr{C})}{H(\mathscr{C})} > I_{max} \tag{7}$$

An alternative form of the stopping criterion in Eq. (7) is based on the normalized mutual information gain being less than a specified positive scalar threshold η_{stop} as stated below:

$$I(\mathscr{P}(\Lambda_{|\Sigma|}^*); \mathscr{C}) - I(\mathscr{P}(\Lambda_{|\Sigma|-1}^*); \mathscr{C}) \leq \eta_{stop} \tag{8}$$

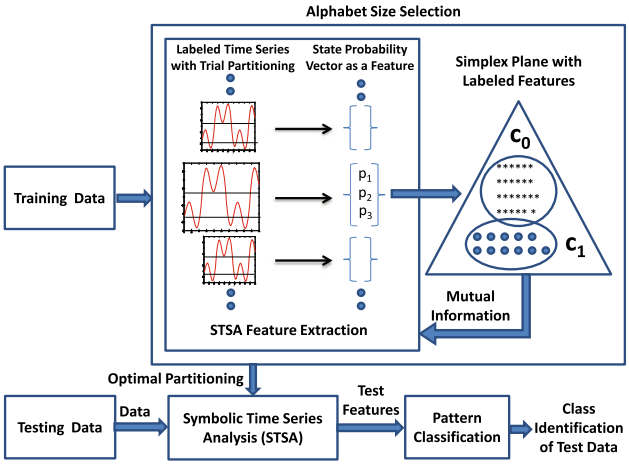

Fig. 1 Information-theoretic framework for time series partitioning and alphabet size selection

In contrast to the direct search of the entire partitioning space, the computational complexity in the current setting increases linearly with $|\Sigma|$. Thus, the proposed approach would allow a finer grid size for the partitioning search with relatively low computational complexity. Figure 1 elucidates an outline of the alphabet size selection method explained above for $|\mathscr{C}| = 2$.

Robustness: This subsection addresses robustness of classification performance when perturbed in partition locations. In this procedure, a zero-mean Gaussian noise is added to generate samples of random boundary locations. If a large number of samples are drawn, the effect of the perturbations is realized from the statistical characteristics of the set of mutual information values corresponding to each sample. To this end, ith partition boundary location is drawn as samples from a Gaussian distribution $N(\lambda_i, \sigma_i)$ (i.e., with mean λ_i and standard deviation σ_i). In particular, σ_i's are chosen to be fractions of $\min(\lambda_i - \lambda_{i-1}, \lambda_{i+1} - \lambda_i)$. At each step, M samples are drawn from the distribution centered at a partition location λ_i that is not yet included in the partition set, i.e., the jth independent and identically distributed (iid) sample ${\lambda_i}^j \sim N(\lambda_i, \sigma_i)$, $j = 1, \ldots, M$. The mutual information corresponding to the features extracted after incorporating ${\lambda_i}^j$ into the existing partition set is obtained as $(I(\mathscr{P}(\Lambda_i^j); \mathscr{C})$, where $\Lambda_i^j = \Lambda_{i-1}^* \cup \{\lambda_i^j\}$.

The mutual information of the feature vectors resulting from adding λ_i into the existing partition set in this paper as 95th percentile of the set of mutual information values corresponding to M samples drawn from the distribution centered at λ_i, i.e.,

$$I(\mathscr{P}(\Lambda_i); \mathscr{C}) = P_{95}\{I(\mathscr{P}(\Lambda_i^j); \mathscr{C}),\ j = 1, \ldots, M\} \quad (9)$$

Hence, at the ith step, the (suboptimal) partition location λ_i^* is obtained as:

$$\lambda_i^* = \arg\max_{\lambda_i}((I(\mathscr{P}(\Lambda_i); \mathscr{C}))) \quad (10)$$

Parzen Window based Mutual Information: This subsection explains usage of the Parzen window method [19,20] for estimation of mutual information at each step of sequential optimization. In classification problems, a class has discrete values, while the input features $\mathscr{P}$ (i.e., state probability vectors $\mathbf{p}$ of PFSA) are usually continuously varying. Similar to its usage in Eq. (3), the mutual information between input feature $\mathscr{P}$ and class $\mathscr{C}$ becomes:

$$I(\mathscr{P}; \mathscr{C}) = H(\mathscr{C}) - H(\mathscr{C}|\mathscr{P}) \quad (11)$$

where $H(\mathscr{C})$ is obtained with a uniform class prior. The conditional entropy $H(\mathscr{C}|\mathscr{P})$ based on the input feature $\mathscr{P}$ (that is a $|Q|$-dimensional random vector) is obtained as:

$$H(\mathscr{C}|\mathscr{P}) = -\int_{\mathscr{P}} p_{\mathscr{P}}(\mathbf{p}) \times \sum_{i=0}^{|\mathscr{C}|-1} p_{\mathscr{C}|\mathscr{P}}(c_i|\mathbf{p}) \log_2 p_{\mathscr{C}|\mathscr{P}}(c_i|\mathbf{p})\, d\mathbf{p} \quad (12)$$

Via applying the Bayesian rule and $\sum_{i=0}^{|\mathscr{C}|-1} p_{\mathscr{C}|\mathscr{P}}(c_i|\mathbf{p}) = 1$ for any given $\mathbf{p}$, the probability $p_{\mathscr{C}|\mathscr{P}}(c|\mathbf{p})$ is becomes:

$$p_{\mathscr{C}|\mathscr{P}}(c|\mathbf{p}) = \frac{p_{\mathscr{P}|\mathscr{C}}(\mathbf{p}|c)\, p_{\mathscr{C}}(c)}{\sum_{i=0}^{|\mathscr{C}|-1} p_{\mathscr{P}|\mathscr{C}}(\mathbf{p}|c_i)\, p_{\mathscr{C}}(c_i)} \quad (13)$$

The Parzen window estimator at each class $c \in \mathscr{C}$ is obtained as:

$$\hat{p}_{\mathscr{P}|\mathscr{C}}(\mathbf{p}|c) = \frac{1}{n_c} \sum_{i \in \mathscr{I}_c} \varphi(\mathbf{p} - \mathbf{p}^i, h_c) \quad (14)$$

where n_c is the number of training samples belonging to the class $c \in \mathscr{C}$ and $\mathscr{I}_c$ is the set of the respective indices of training samples (i.e., $|\mathscr{I}_c| = n_c$); φ is the Parzen window function; and h_c is the Parzen window width parameter for the pattern class $c \in \mathscr{C}$. In this paper, a d-variate Gaussian window (with covariance matrix S) is chosen as the Parzen density estimator.

$$\varphi(\mathbf{p}, h) = \frac{1}{(2\pi)^{d/2} h^d |S|^{1/2}} \exp\left(-\frac{\mathbf{p}^T S^{-1} \mathbf{p}}{2h^2}\right) \quad (15)$$

where the parameter h controls the tradeoff between variance and bias of the estimator. An increment in h would reduce the variance at the expense of increased bias and vice-versa for a decrement in h. Following [19], the current paper uses $h_c = \frac{1}{2\log_e(n_c)}$ for each $c \in \mathscr{C}$. Parzen [20] showed that the estimated density converges to the true density if φ

and h are selected properly. By combining Eqs. (13), (14) and (15), the Parzen window estimator is constructed [19] as:

$$\hat{p}_{\mathscr{C}|\mathscr{P}}(c|\mathbf{p}) = \frac{\sum_{i\in\mathscr{I}_c} \exp\left(-\frac{(\mathbf{p}-\mathbf{p}^i)^T S^{-1}(\mathbf{p}-\mathbf{p}^i)}{2h^2}\right)}{\sum_{k=0}^{|\mathscr{C}|-1}\sum_{j\in\mathscr{I}_k} \exp\left(-\frac{(\mathbf{p}-\mathbf{p}^j)^T S^{-1}(\mathbf{p}-\mathbf{p}^j)}{2h^2}\right)} \tag{16}$$

If the integration in Eq. (12) is replaced by summation of the sample points with equal sample probability, then the conditional entropy (based on an input feature $\mathscr{P}$) derived from the training data belonging to all classes in $\mathscr{C}$ becomes:

$$\hat{H}(\mathscr{C}|\mathscr{P}) = -\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{|\mathscr{C}|-1} \hat{p}_{\mathscr{C}|\mathscr{P}}(c_i|\mathbf{p}^j)\log_2 \hat{p}_{\mathscr{C}|\mathscr{P}}(c_i|\mathbf{p}^j) \tag{17}$$

where $n \triangleq \sum_{i=0}^{|\mathscr{C}|-1} n_{c_i}$ is the total number of training samples under consideration and $\mathbf{p}^j$ is the feature vector computed from the jth training data in the ensemble of all classes. Finally, the estimated mutual information is obtained from Eqs. (16) and (17). For $|\mathscr{C}| \ll n$, the computational complexity of Parzen window estimation [19] in Eq. (17) is of the order $n^2 \times d$; this implies that, unlike the histogram-based methods, Parzen window estimation does not require excessive memory.

Pattern Classification using STSA Features: The sub-optimal partitioning obtained by the above-mentioned procedure is used to construct a PFSA from each training and testing time series ; the state probability vector of the PFSA is the extracted feature for each time series. In this paper, two commonly used pattern classifiers, namely, k nearest neighbor (k-NN) and support vector machine (SVM) have been adopted [21].

4 Algorithm validation and results

This section presents two examples for validation of the proposed procedure.

Example #1: Duffing System Simulation: The exogenously excited Duffing system is nonlinear and exhibits complex behavior with chaotic and bifurcation properties; its governing equation is:

$$\frac{d^2y}{dt^2} + \beta\frac{dy}{dt} + \alpha y(t) + y^3(t) = A\,cos(\omega t) \tag{18}$$

where the amplitude A = 22.0, excitation frequency $\omega = 5.0$, and the initial conditions are: $y(0) = 1.0$ and $\frac{dy}{dt}(0) = 0.0$; however, these initial conditions have no significance because only the steady-state oscillatory responses have been analyzed. At first, only two classes of Duffing system are defined based on the range of β, that are: (i) Class 1 ($0.100 \leq \beta \leq 0.147$) and (ii) Class 2 ($0.147 \leq \beta \leq 0.194$). Two hundred simulation runs of the Duffing system have been conducted for each class to generate data set for analysis among which 30 samples are chosen for determining the optimal partitioning, and three-fold cross validation has been performed on the remaining data set to determine the classification results. Parameters α and β are chosen randomly from independent uniform distributions within the prescribed ranges. The length of the simulation time window is 80 s sampled at 100 Hz, which generates 8000 data points. The range of time series is divided into 40 grid cells via Uniform Partitioning.

The proposed sequential partitioning optimization procedure is then employed to identify the optimal partitioning and alphabet size. Figure 2a, b depict the nature of mutual information between the state probability vector and the class labels for the depth of the D-Markov machine of the input feature being $D = 1$ and $D = 2$, respectively. For $D = 2$, normalized mutual information converges to 1 much earlier for alphabet size $|\Sigma| = 5$ than that for $D = 1$ and $|\Sigma| = 9$. In each case, stopping criterion follows Eq. (8) with the parameter $\eta_{stop} = 0.01$. Figure 3a–c show the classification performance of the k-NN classifier [21] with $k = 5$ for three different levels of robustness, i.e., different variances that are fractions, 0.67, 0.5, 0.25, of the inter partition width, respectively. It is observed that the classification errors are smaller with smaller variance which is a consequence of smaller robustness variance fractions. Figure 2c shows the nature of mutual information between the state probability vector and the class labels for the Duffing system with 4 classes, corresponding to four different combinations of the ranges: $\big((0.100 \leq \beta \leq 0.147), (0.147 \leq \beta \leq 0.194), (0.934 \leq \alpha \leq 1.067), (0.8 \leq \alpha \leq 0.934)\big)$, within which the parameters α and β in Eq. (18) are located. The convergence rate of the normalized mutual information is smaller in this case than that for the binary classification scheme because a larger alphabet is required to capture the information of four classes. The stopping criterion follows Eq. (8) with the parameter $\eta_{stop} = 0.01$.

Example #2: LBO Prediction in a Combustor: Ultra-lean combustion is commonly used for NOx reduction and is susceptible to thermo-acoustic instabilities and lean blowout (LBO) [22]. It is well known that occurrence of LBO could be detrimental for operations of both land-based and aircraft gas turbine engines. In essence, a sudden decrease in the equivalence ratio may lead to LBO in gas turbine engines, which could have serious consequences. This event calls for early detection and accurate prediction of LBO for adequate control.

The proposed procedure of time series partitioning and alphabet size selection has been evaluated under multiple

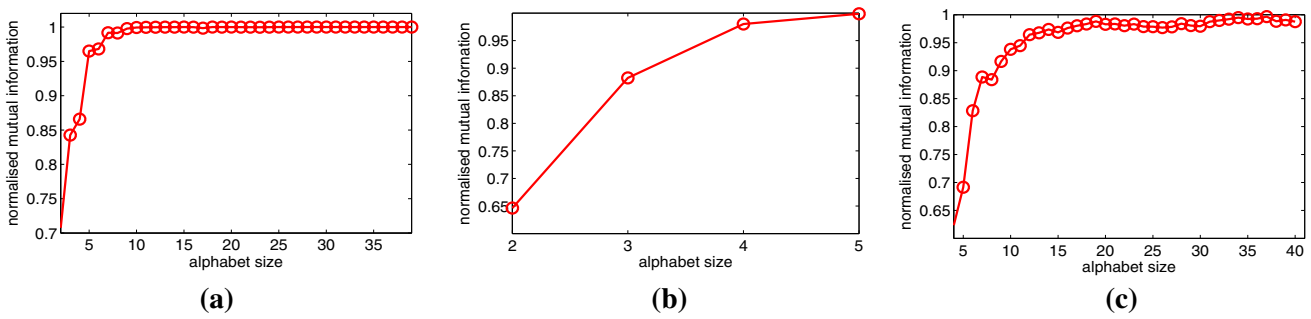

Fig. 2 Mutual information as a function of alphabet size $|\Sigma|$ for two-class Duffing system with **a** $D = 1$, **b** $D = 2$; for **c** four-class Duffing system with $D = 1$ and $\eta_{stop} = 0.01$

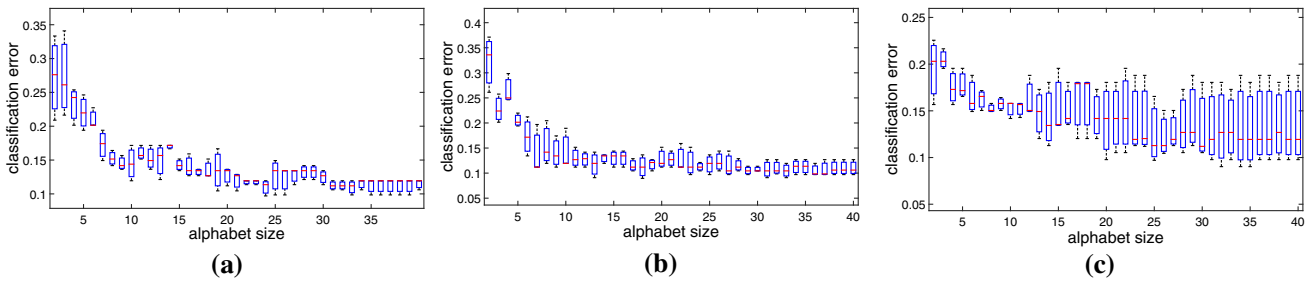

Fig. 3 Misclassification error for two-class problem ($D = 1$) with variance fraction: **a** 0.67, **b** 0.5, **c** 0.25

operating conditions (e.g., airflow rates and premixing levels of fuel and air) on a laboratory apparatus, a detailed description of which is reported in [22]. A series of experiments have been conducted on this laboratory apparatus with liquefied petroleum gas (LPG) fuel at airflow rates of 150, 175 and 200 liters per minute (lpm) for two different fuel-air premixing lengths (i.e., distance of fuel injection port from the dump plane) of L_{fuel} = 25 cm, and 15 cm for Port 3, and Port 5, respectively [22]. For each experiment protocol, chemiluminescence time series data were collected while reducing the fuel-air ratio ϕ in steps till the combustor system reached LBO. The main challenge here is to predict quantitatively how far a combustion process is from the onset of LBO in real time. It is easier to predict LBO under high premixing (i.e., port 3) as the precursor events are more dominant [22] than that under lower premixing (i.e., port 5).

A nested classification architecture [22] is proposed in accordance with the range of the non-dimensional equivalence ratio of ϕ/ϕ_{LBO} for early detection of LBO. In the training phase, the chemiluminescence time series of duration $3\,s$ (at the sampling rate of 2 kHz) for each premixing length are grouped into two classes as: *Alarm* ($1 \leq \phi/\phi_{LBO} \leq 1.20$) and *Nominal* ($\phi/\phi_{LBO} > 1.20$). The class *Alarm* is subdivided into two finer classes as: *Impending LBO* (ILBO) for $1 \leq \phi/\phi_{LBO} \leq 1.1$, and *Progressive LBO* (PLBO) for $1.1 < \phi/\phi_{LBO} \leq 1.2$. Identification of the PLBO phase is critical for avoidance of LBO as control actions need to be initiated typically near the PLBO-ILBO boundary. The proposed sequential partitioning optimization scheme starts with 20 grid cells. The robustness is chosen as σ fraction of 0.25. Figure 4a shows the variations of mutual information between the D-Markov feature vectors with $D = 1$ and the class labels for both premixing levels; Figure 4b presents a similar analysis for $D = 2$. It is observed from these results that the normalized mutual information converges to 1 with a much smaller alphabet size $|\Sigma|$ for $D = 2$ than that for $D = 1$, reflecting the fact that D-Markov features with larger memory should be able to capture the same class information with a smaller $|\Sigma|$; however, the number of PFSA states for $D = 2$ could be larger than that for $D = 1$. Normalized mutual information for high premixing (port 3) converges to 1 for a smaller $|\Sigma|$ than that for low premixing (port 5). This phenomenon is more apparent for $D = 1$, where the alphabet size for port 3 and port 5 is chosen as $|\Sigma| = 7$ and $|\Sigma| = 12$, respectively, according to a stopping rule of $I_{max} = 1$ (see Eq. (7)). The rationale for this observation is attributed to large class separability for high premixing due to the presence of dominant precursor events leading to LBO. Support vector machines (SVM) with radial basis functions [21] have been used based on 70 % training at each layer of the nested classification. Variance of the radial basis function is optimized for each layer of the nested classification via a grid search method, and it is found to be 1 in most of the cases. Figure 4c, d present the variations of classification error, while the proposed partitioning scheme sequentially increases $|\Sigma|$ for both $D = 1$ and $D = 2$. The

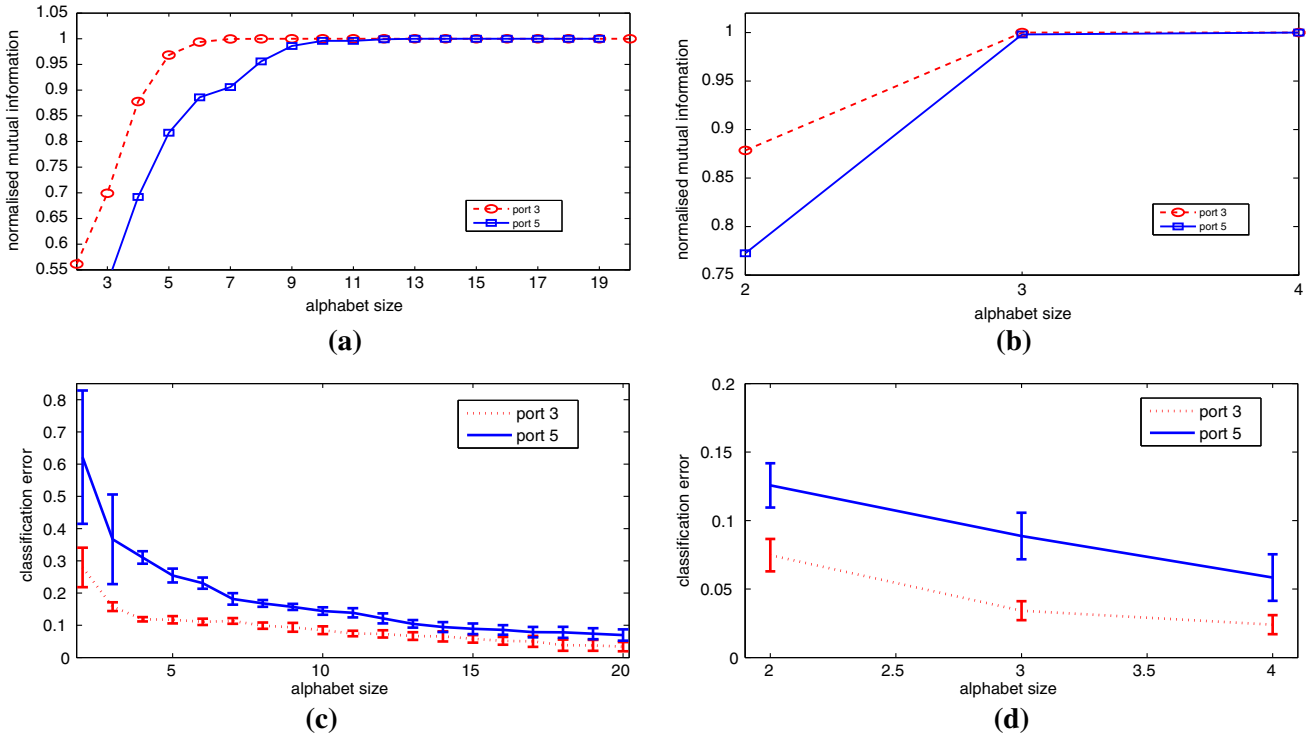

Fig. 4 **Top**: Mutual information as a function of alphabet size $|\Sigma|$ for Port 3 and Port 5 levels of premixing with **a** $D = 1$ and **b** $D = 2$ **Bottom**: variation of classification error as a function of alphabet size $|\Sigma|$ for Port 3 and Port 5 levels of premixing level with **c** $D = 1$ and **d** $D = 2$

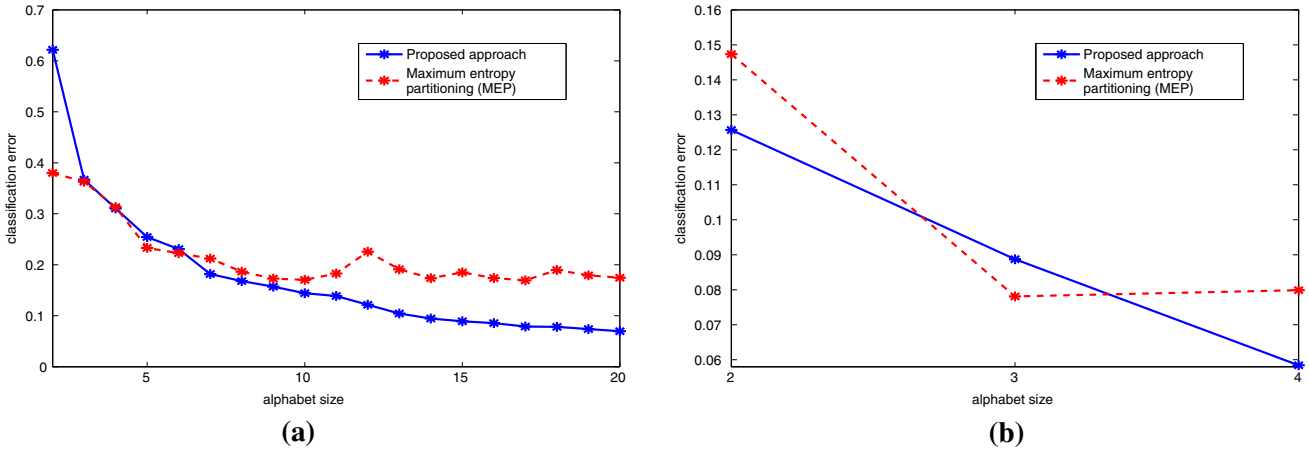

Fig. 5 Comparison of classification error with varying alphabet size for (Port 5) premixing: **a** $D = 1$, **b** $D = 2$

error bars correspond to standard deviations of the classification error over 10-fold cross validation. The classification error is smaller for high premixing (port 3) than that for low premixing (port 5). It is also observed that relatively smaller classification error occurs at $D = 2$ than that at $D = 1$, especially for small $|\Sigma|$.

Performance Comparison: Performance of the proposed partitioning is compared with that of a benchmark partitioning method, namely, maximum entropy partitioning (MEP). Figure 5 shows the profiles of classification error in the proposed approach and in MEP as a function of $|\Sigma|$ for the port 5 scenario. By applying normalized mutual information stopping criterion as mentioned earlier, the classification error is seen to be smaller for the proposed scheme at $D = 1$ and $|\Sigma| \geq 6$ in Figure 5a and at $D = 2$ and $|\Sigma| = 4$ in Figure 5b.

5 Summary and conclusions

This paper addresses the issues of: (i) alphabet size selection and partitioning of time series data for symbolization of time series, and (ii) information-theoretic analysis with a focus on

feature extraction and pattern classification from sensor data in dynamic data-driven application systems (DDDAS) [9]. The feature extraction algorithm maximizes the mutual information between the input features and pattern classes in the framework of symbolic time series analysis (STSA) [2]. The proposed technique is validated on two examples: (i)simulation data for an exogenously excited Duffing system [23] and (ii) experimental data of chemiluminescence time series generated from a swirl-stabilized combustor [22] for lean blowout (LBO) prediction.

The proposed partitioning technique yields satisfactory performance of pattern classification in several test phases. Nevertheless, its efficacy may depend on the very nature of the time series under consideration. Incorporation of an explicit term for class separability in the proposed objective function is a topic of future investigation. Apart from this issue, the following research topics are recommended for future research: (i) implementation of simultaneous optimization techniques instead of sequential ones, (ii) tradeoff between the performance gain and the loss of computational speed, and (iii) validation of the proposed algorithm for other applications of pattern classification.

Acknowledgments The work reported in this paper has been supported in part by the US Air Force Office of Scientific Research (AFOSR) under Grant No. FA9550-15-1-0400.

References

1. Beim Graben, P.: Estimating and improving the signal-to-noise ratio of time series by symbolic dynamics. Phys. Rev. E **64**(5), 051104 (2001)
2. Daw, C., Fenney, C., Tracy, E.: A review of symbolic analysis of experimental data. Rev. Sci. Instrum. **74**, 915–930 (2003)
3. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: a novel symbolic representation of time series. Data Min. Knowl. Discov. (2007). doi:10.1007/s10618-007-0064-z
4. Lind, D., Marcus, B.: An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge (1995)
5. Ray, A.: Symbolic dynamic analysis of complex systems for anomaly detection. Signal Process. **84**(7), 1115–1130 (2004)
6. Rajagopalan, V., Ray, A.: Symbolic time series analysis via wavelet-based partitioning. Signal Process. **86**(11), 3309–3320 (2006)
7. Subbu, A., Ray, A.: Space partitioning via Hilbert transform for symbolic time series analysis. Appl. Phys. Lett. **92**(8), 084107 (2008)
8. Mukherjee, K., Ray, A.: State splitting and merging in probabilistic finite state automata for signal representation and analysis. Signal Process. **104**, 105–119 (2014)
9. Darema, F.: Dynamic data driven applications systems: new capabilities for application simulations and measurements. In: 5th International Conference on Computational Science - ICCS 2005, (Atlanta, GA, USA), (2005)
10. Rao, C., Ray, A., Sarkar, S., Yasar, M.: Review and comparative evaluation of symbolic dynamic filtering for detection of anomaly patterns. Signal Image Video Process. **3**(2), 101–114 (2009)
11. Bahrampour, S., Ray, A., Sarkar, S., Damarla, T., Nasrabadi, N.: Performance comparison of feature extraction algorithms for target detection and classification. Pattern Recognt. Lett. **34**, 2126–2134 (2013)
12. Dupont, P., Denis, F., Esposito, Y.: Links between probabilistic automata and hidden Markov models: probability distributions, learning models and induction algorithms. Pattern Recognit. **38**(9), 1349–1371 (2005)
13. Buhl, M., Kennel, M.: Statistically relaxing to generating partitions for observed time-series data. Phys. Rev. E **71**(4), 046213 (2005)
14. Sarkar, S., Mukherjee, K., Jin, X., Singh, D., Ray, A.: Optimization of symbolic feature extraction for pattern classification. Signal Process. **92**(3), 625–635 (2012)
15. Sarkar, S., Chattopadhyay, P., Ray, A., Phoha, S., Levi, M.: Alphabet size selection for symbolization of dynamic data-driven systems: an information-theoretic approach. In: 2015 American Control Conference (ACC), (Chicago, OH, USA), pp. 5194–5199, July 1–3 (2015)
16. Cover, T., Thomas, J.: Elements of Information Theory, 2nd edn. Wiley, Hoboken, NJ, USA (2006)
17. Steuer, R., Molgedey, L., Ebeling, W., Jimenez-Montano, M.: Entropy and optimal partition for data analysis. Eur. Phys. J. B **19**, 265–269 (2001)
18. Jin, X., Gupta, S., Mukherjee, K., Ray, A.: Wavelet-based feature extraction using probabilistic finite state automata for pattern classification. Pattern Recognit. **44**(7), 1343–1356 (2011)
19. Kwak, N., Choi, C.: Input feature selection by mutual information based on parzen window. IEEE Trans. Pattern Anal. Mach. Learn. **24**(12), 1667–1671 (2002)
20. Parzen, E.: On estimation of a probability density function and mode. Ann. Math. Stat. **33**, 1065–1076 (1962)
21. Bishop, C.M.: Pattern Recognit. Mach. Learn. Springer, New York (2006)
22. Sarkar, S., Ray, A., Mukhopadhyay, A., Sen, S.: Dynamic data-driven prediction of lean blowout in a swirl-stabilized combustor. Int. J. Spray Combust. Dyn. **7**(3), 209–242 (2015)
23. Thompson, J., Stewart, H.: Nonlinear Dynamics and Chaos. Wiley, Chichester (1986)