On Compression of Machine-derived Context Sets for Fusion of Multi-modal Sensor Data*

Nurali Virani, Shashi Phoha, and Asok Ray

Abstract Dynamic data-driven application systems (DDDAS) operate on a sensing infrastructure for multi-modal measurement, communications, and computation, through which they perceive and control the evolution of physical dynamic processes. Sensors of different modalities are subject to contextually variable performance under varying operational conditions. Unsupervised learning algorithms have been recently developed to extract the operational context set from multi-modal sensor data. This context set represents the set of all natural or man-made factors, which along with the state of the system, completely condition the measurements from sensors observing the system. The desirable property of conditional independence of observations given the state-context pair enables tractable fusion of disparate information sources. In this chapter, we address a crucial problem associated with unsupervised context learning of reducing the cardinality of the context set. Since, the machine-derived context set can have a large number of elements, we propose one graph-theoretic approach and one subset selection approach for the controlled reduction of contexts to obtain a context set of lower cardinality. We also derive an upper bound on the error introduced by the compression. These proposed approaches are validated with data collected in field experiments with unattended ground sensors for border-crossing target classification.

Nurali Virani and Asok Ray
Department of Mechanical and Nuclear Engineering, The Pennsylvania State University, University Park, PA 16802, USA. Nurali Virani is now with GE Global Research, Niskayuna, NY 12309, USA. e-mail: nurali.virani88@gmail.com, axr2@psu.edu

Shashi Phoha
Applied Research Laboratory, The Pennsylvania State University, University Park, PA 16802, USA. e-mail: sxp26@arl.psu.edu

*The work reported in this chapter has been supported in part by U.S. Air Force Office of Scientific Research (AFOSR) under Grant No. FA9550-12-1-0270. Any opinions, findings, and conclusions in this chapter are those of the authors and do not necessarily reflect the views of the sponsoring agencies.

1 Introduction

Dynamic data-driven application systems (DDDAS) rely on information from a multitude of sensors to assess the state of any observed system [1]. This system can not only observe and control states of a physical system, but also adapt the sensing system to obtain better understanding of the system. It is well known that the measurements from the (possibly) multi-modal sources of information are affected not only by the state, but also by the operational conditions around the system [2]. These natural or man-made factors are known as context in literature [3, 4, 5, 6]. For example, soil moisture, soil porosity, and ground temperature are contexts for seismic sensors, whereas wind speed and air temperature are those for acoustic sensors. Physics-based analytical models try to capture some of the contextual effects in great detail, but they need accurate estimates of several time-varying environmental parameters. On the other hand, there have been efforts to develop data-driven models for unsupervised discovery of contexts from sensor data. The complexity and accuracy of context-aware DDDAS in state estimation and measurement system adaptation is directly affected by the size of the context set obtained from data-driven or physics-based techniques. This chapter will focus on unsupervised data-driven learning of context with a specific emphasis on techniques to compress the set of contexts and also to understand the effect of this compression. The compression enables to ensure real-time implementation of DDDAS without significantly affecting the performance, such as accuracy of the system.

Recently, in [5] the notion of *context* was mathematically formalized to enable machines to learn from data and then use context in decision-making. However in [3], context was defined as a parameter which along with user-defined state set of the system completely conditions the measurements. Unlike several existing contextual reasoning frameworks, where context is usually associated with a specific modality, proposing context as the enabler of conditional independence unifies the notion of context across all modalities in the system. Bayesian fusion uses the likelihood of a new measurement given all previous measurements from other sensors to correctly obtain the posterior density. This process becomes intractable for systems which have more than a couple of sensors. Thus, context enables tractable fusion of multi-modal sensors without relying on the possibly incorrect assumption of conditional independence given only the system state (Naïve Bayes assumption). Unsupervised learning of context using clustering as well as density-estimation based approaches have been reported in literature. These approaches either focused on single sensor systems, such as ground penetrating radars in [4] or video sensors in [7], or it is just assumed that the machine-derived context would provide conditional independence in measurements given the state-context pair [5]. However, the non-parametric density-estimation approach for context learning in [3] guarantees that given any user-defined state and machine-defined context pair, the conditional independence property holds true.

The size of context set directly affects the memory required and computation time of the context-aware decision-making approaches for sensor selection [8], tracking [9], multi-modal fusion [6], and pattern recognition [5]. In wireless sen-

sor network applications, such as border surveillance, where power, memory, and execution time are severely constrained, we need to be able to restrict the size of context sets to enable tractable execution of context-aware approaches on resource-constrained platforms. Thus, in this work, we explore different approaches for context set compression. The context learning approach in [3] relies on a convex optimization formulation using the concept of kernel-based density estimation [10, 11]. Thus, adding any explicit sparsity constraint makes the problem non-convex and hard to solve. Moreover, enforcing strict sparsity constraint on the solution can severely affect the performance of the solution as model order and accuracy are known to be competing objectives [12] and one might have to repeat the non-convex optimization several times before obtaining a solution with acceptable error. This motivates the need for augmenting the original *convex optimization formulation* with a separate compression step in which the additional maximum error incurred is directly controlled.

This chapter will first review some important aspects of the original optimization problem in Section 2. The main objective of the chapter is to introduce two distinct techniques to compress the set of contexts and quantify the effect of this compression on the accuracy of the density estimate. The first proposed technique uses the classical graph-theoretic problem of *maximal clique enumeration* [13] for compression of context sets by using a depth-first search strategy [14]. The second technique presents a *subset-selection approach* and establishes a relation of the compression ratio with the upper bound on the additional error incurred by compression on the density estimate. These techniques are explained in Section 3. The techniques developed in this work are validated using data collected in field experiments from a border surveillance testbed with two geophones to classify whether a human target is walking or running. Finally in Section 5, the concluding remarks are presented.

2 Learning Context from Data

Data-driven modeling of context has been only recently explored in the field of machine learning to enhance the process of information fusion. Unsupervised learning methods using k-means and modularity-based clustering have been reported for obtaining modality-specific context sets [5]. Density estimation has also been used for learning context from data. A parametric approach for context learning to obtain Gaussian mixture models was presented in [4], whereas a nonparametric approach using kernel-based regression was proposed in [3]. In this section, we first present a definition of context, which mathematically formalizes this widely-used notion of context, and then we review some existing methods to derive context sets from data.

Definition 2.1. (Context and Context Set [3]) Suppose that the measurements Y_1 and Y_2 take values in $\mathscr{Y}_1$ and $\mathscr{Y}_2$, respectively. Suppose that the state X takes values from a finite set $\mathscr{X}$. Then, a nonempty finite set $\mathscr{C}(X)$ is called the *context set* and each element $c \in \mathscr{C}(X)$ of the set is called a *context*, if the measurements Y_1 and Y_2

are mutually independent conditioned on the state-context pair (x,c) for all $x \in \mathscr{X}$ and for all $c \in \mathscr{C}(X)$.

According to this definition, the following relation holds:

$$p(Y_1,Y_2 \mid X,c) = p_1(Y_1 \mid X,c)p_2(Y_2 \mid X,c) \quad \text{for all } c \in \mathscr{C}(X). \tag{1}$$

Here, the left-hand side of (1) denotes the conditional density of (Y_1,Y_2) given (X,c), and the right-hand side gives the product of conditional densities of Y_1 and Y_2 given (X,c). The Definition 2.1 enables to obtain a single context set from multi-modal sensor data, when the measurement space $\mathscr{Y}_1$ and $\mathscr{Y}_2$ corresponds to heterogeneous sensors. In order to generate a context set $\mathscr{C}(x)$ for each $x \in \mathscr{X}$, so that (1) holds, kernel-based density estimation [10] was used in [3]. This context learning process is outlined ahead in this section.

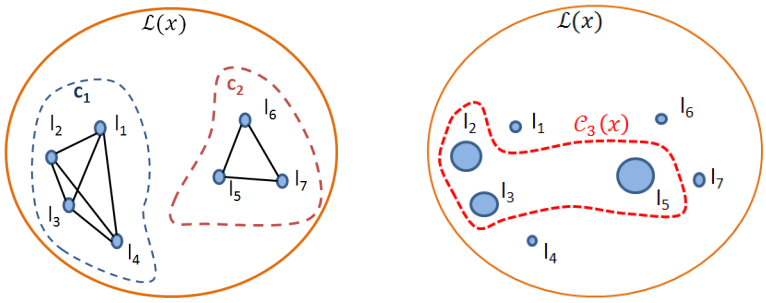
The problem of obtaining all contexts, which satisfy the relation in (1), is non-trivial and the concept to pose it as a nonparametric mixture modeling problem was first proposed in [3]. In view of Definition 2.1, the measurement likelihood function is of the form

$$\begin{aligned} p(Y_1,Y_2 \mid X) &= \sum_{c \in \mathscr{C}(X)} \pi_c(X) p(Y_1,Y_2 \mid X,c) \\ &= \sum_{c \in \mathscr{C}(X)} \pi_c(X) p_1(Y_1 \mid X,c) p_2(Y_2 \mid X,c), \end{aligned} \tag{2}$$

where $\pi_c(X)$ is the prior probability that, conditioned on the state X, the true context is c. In order to estimate this likelihood model, the conditional density was represented as the following mixture model

$$p(Y_1,Y_2 \mid X) = \sum_{c \in \mathscr{C}(X)} \pi_c(X) K_1\big(s_1^{(c)}(X),Y_1\big) K_2\big(s_2^{(c)}(X),Y_2\big), \tag{3}$$

where the prior probability $\pi_c(X)$ denotes the weight of the component corresponding to context c in the context set $\mathscr{C}(X)$ and the component is represented by the product of kernel functions $K_i \colon \mathscr{Y}_i \times \mathscr{Y}_i \to \mathbb{R}$ for $i = 1,2$. Also, $s_i^{(c)}(X) \in \mathscr{Y}_i$ is a support vector [10] obtained by solving the kernel regression problem using training data consisting of the triples (Y_1,Y_2,X). Thus, the problem of learning context set was reduced to that of identifying support vectors of a regression problem. The details of this technique are available in [3]. The context set identified by regression has error bounded above by an insensitivity parameter, which is chosen by the user. Although, one can use this error margin parameter to indirectly influence the size of context set, there is no explicit relationship for the set cardinality with the chosen error margin. Thus, the main contribution of the chapter is explained in the next section, which enables controlled compression of context sets.



(a) Maximal clique enumeration (b) Subset selection

Fig. 1: Concept of context set compression

Algorithm 1: Context Set Compression by Maximal Clique Enumeration

Input: Observation densities $p(Y \mid X, L)$ and threshold ε.
Output: Context set $\mathscr{C}(X)$.

```
for all x ∈ 𝒳 do
    Compute weight matrix W(x);
    𝒢_{x,ε} = ConstructGraph(W(x), ε);
    ℳ = MCE(𝒢_{x,ε});
    𝒞(x) = Minterms(ℳ).
```

3 Cardinality Reduction of Context Sets

This section will explain the two proposed techniques for cardinality reduction of context sets using the maximal clique enumeration algorithm from graph theory and a simple subset selection approach. These techniques assume that the density estimation step for unsupervised context learning has already been solved and the resulting density estimate is used in both of these techniques.

3.1 Graph-Theoretic Compression

In graph theory, a clique is a complete subgraph and it is maximal, if it is not contained in a bigger clique. Maximal clique enumeration (MCE) is a classical problem in graph theory, which was addressed in detail in [13, 14] using depth-first search strategy. We use this concept to identify all machine-derived contexts whose effect on sensor measurements is almost identical. The context set is used as the vertex set of a weighted graph and the edge weights denote the pairwise distance between contextual observation densities. The MCE-based context set compression approach is explained in detail ahead in this section.

Let $l_1, l_2, \ldots, l_{|\mathscr{L}(x)|}$ denote the distinct machine-derived contexts for the state $x \in \mathscr{X}$ before compression and let c denote an element of the compressed context set $\mathscr{C}(x)$ for the state $x \in \mathscr{X}$. The algorithm defines a weight matrix $\mathbf{W}(x) = [w_{ij}(x)] \in \mathbb{R}^{|\mathscr{L}(x)| \times |\mathscr{L}(x)|}$ by

$$w_{ij}(x) = d\big(p(Y \mid X = x, L = l_i), p(Y \mid X = x, L = l_j)\big)$$

for $x \in \mathscr{X}$ and $i, j = 1, \ldots, |\mathscr{L}(x)|$, where $Y = (Y_1, Y_2, \ldots, Y_N)$ is the concatenated measurement from all sensors. Here, denoted by $d(\cdot,\cdot)$ is a distance function on the space of observation densities, such as symmetric Kullback-Leibler divergence [15] or the Bhattacharyya distance [16]. For a chosen positive real number ε, let $\mathscr{G}_{x,\varepsilon}$ denote the ε-context graph for state $x \in \mathscr{X}$, which is defined by the tuple $(\mathscr{L}(x), \mathscr{E}(x,\varepsilon))$, where the vertex set $\mathscr{L}(x)$ represents the set of all machine-derived contexts corresponding to the state $x \in \mathscr{X}$ and the edge set is given as

$$\mathscr{E}(x,\varepsilon) = \{(l_i, l_j) \in \mathscr{L}(x)^2 \colon w_{ij}(x) \leq \varepsilon, i, j = 1, \ldots, |\mathscr{L}(x)|\}$$

for each $x \in \mathscr{X}$. This graph $\mathscr{G}_{x,\varepsilon}$ is constructed in the `ConstructGraph` function. The edge set $\mathscr{E}(x,\varepsilon)$ represents all pairs of context whose measurement densities are at most distance ε away from each other. This graph $\mathscr{G}_{x,\varepsilon}$ is then processed by the `Maximal Clique Enumeration` function, which implements the depth-first search strategy given in [14], to obtain the set of all maximal cliques denoted by $\mathscr{M}$. Each maximal clique is a subset of the context set consisting of contexts which are all mutually at most distance ε away from each other. The maximal cliques will form a *set cover* of the set $\mathscr{L}(x)$ (i.e., union of maximal cliques equals the complete set), but they can end up being overlapping, thus, denoting each clique as a context can lead to the loss of the desired conditional independence property. Moreover, it is known that for an n-vertex graph, the maximum number of maximal cliques is given by $3^{n/3}$ [17], thus, the resulting context set might become exponentially larger. Hence, the function `Minterms`$(\mathscr{M})$ uses the method in [18] to evaluate all minterms of $\mathscr{M}$ (i.e., nonempty set differences and intersections formed by the members of $\mathscr{M}$) to obtain a mutually exclusive and exhaustive collection $\mathscr{C}(x)$ of cliques that partition the set $\mathscr{L}(x)$; for example, `Minterms`$(\{\{1,2,3,5\},\{2,4\}\})$ gives $\{\{1,3,5\},\{2\},\{4\}\}$. These steps are given in the Algorithm 1, which lead to the construction of the compressed context set denoted by $\mathscr{C}(X)$.

Each element $c \in \mathscr{C}(X)$ of the context set is a collection of the machine-defined contexts $l \in \mathscr{L}(X)$. The corresponding contextual observation density and prior distribution need to be derived for the compressed context set. We will first assign values to $p(c \mid X, l)$ as follows:

$$p(c \mid X, l) = \begin{cases} 1, & \text{if } l \in c \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

This conditional density is well-defined as $\mathscr{C}(X)$ is a partition of $\mathscr{L}(X)$ and it will assume the value of 1 for only one $c \in \mathscr{C}(X)$. Now, we can compute the prior density using (4) as follows:

$$p(c|X) = \sum_{l \in \mathscr{L}(X)} p(c \mid X, l) p(l \mid X) = \sum_{l \in c} p(l \mid X), \tag{5}$$

where $p(l|X)$ is the state-dependent prior probability of the machine-defined context which is known. The observation density can be shown to be given accurately by the mixture model

$$p(Y \mid X, c) = \sum_{l_i \in c} \frac{p(l_i \mid X)}{p(c \mid X)} p(Y \mid X, l_i). \tag{6}$$

The overall model complexity stays the same as the number of mixture components still remain the same. In order to reduce the model complexity, we define the observation density $p(Y \mid X, c) = p(Y \mid X, l^*)$, where, l^* is an element in c. Theorem 3.1 derives an approach to choose l^* and provides the bound for the error induced by this process.

Theorem 3.1 (Bound for error introduced in compression by clique enumeration)**.** *If the distance function d used in Algorithm 1 is symmetric Kullback-Leibler divergence (*sKL*), then for any fixed threshold $\varepsilon > 0$, the error induced by defining $p(Y \mid X, c) = p(Y \mid X, l^*)$ for some $l^* \in c$ is upper-bounded by the value $\varepsilon\left(1 - \frac{p(l^*|X)}{p(c|X)}\right)$, which is strictly less than ε. This error bound is minimized for $l^* = \arg\max_{l \in c} p(l \mid X)$.*

Proof. At first, a known result from literature will be shown and then, we will use it to prove the theorem. Let p_0 denote a mixture model with component densities f_i^0 and weights π_i^0 for $i \in \{1, \ldots, n_0\}$, and similarly p_1 denotes another mixture model with n_1 components. The convexity upper bound on KL-divergence [15] is given by

$$\begin{aligned} \mathrm{KL}(p_0 \parallel p_1) &\leq \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \pi_i^0 \pi_j^1 \, \mathrm{KL}(f_i^0 \parallel f_j^1) \\ \implies d(p_0, p_1) &\leq \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \pi_i^0 \pi_j^1 d(f_i^0, f_j^1), \end{aligned} \tag{7}$$

since, $d(p_0, p_1) = \mathrm{sKL}(p_0, p_1) = \mathrm{KL}(p_0 \parallel p_1) + \mathrm{KL}(p_1 \parallel p_0)$. Let us consider p_0 to be the density from (6) with $\pi_i^0 = \frac{p(l_i|X)}{p(c|X)}$, $f_i^0 = p(Y \mid X, l_i)$, and $n_0 = |c|$, and assign $p_1 = p(Y|X, l^*)$, i.e., $\pi_j^1 = 1$, if $l_j = l^*$, $\pi_j^1 = 0$, otherwise. Using (7) and substituting for the mixture models, we get

$$d(p_0, p_1) \leq \sum_{l_i=1}^{|c|} \pi_i^0 d(p(Y \mid X, l_i), p(Y \mid X, l^*)) = \sum_{l_i \in c \setminus \{l^*\}} \pi_i^0 d(p(Y \mid X, l_i), p(Y \mid X, l^*))$$

using that l^* belongs to c and the positive definiteness property of the distance function. Since we know that ε is the distance threshold and clique c consists only of elements l, whose observations densities are at most ε away from each other, we

obtain $d(p_0, p_1) \leq \sum_{l_i \in c \setminus \{l^*\}} \pi_i^0 \varepsilon$. Substituting the values for π_i^0 and using (5)), we will obtain the desired result as follows:

$$d(p_0, p_1) \leq \varepsilon \sum_{l_i \in c \setminus \{l^*\}} \frac{p(l_i \mid X)}{p(c \mid X)} = \varepsilon \left(1 - \frac{p(l^* \mid X)}{p(c \mid X)}\right). \tag{8}$$

Since l^* belongs to c, we have $p(l^*|X) \leq p(c|X)$. Thus, we verify that $d(p_0, p_1)$ is indeed less than ε and the error bound is minimized for $l^* = \arg\max_{l \in c} p(l \mid X)$. □

This result shows us that, if we choose acceptable level of error (ε), then we can use it as the graph threshold in the function `ConstructGraph` of Algorithm 1. The two limitations of this approach are that: (i) the compression level is not known a priori, and (ii) the computations have to be redone if we decide to change the value of ε. In order to alleviate these two issues, we will look at another approach called subset selection.

3.2 Compression by Subset Selection

In the subset selection approach, we directly choose the size of the desired compressed context set (say k) and not the acceptable error. The proposed approach is to select a set of k distinct contexts from the machine-derived context set $\mathscr{L}(X)$ and assign those to a compressed context set $\mathscr{C}_k(X)$. Thus, we will end up with $\mathscr{C}_k(X) \subsetneq \mathscr{L}(X)$ for $k < |\mathscr{L}(X)|$. This section explains the approach to select the subset and derive a bound on the error introduced by subset selection.

Let us denote $\bar{\mathscr{C}}_k(X)$ as the relative complement of $\mathscr{C}_k(X)$ with respect to $\mathscr{L}(X)$, given by $\bar{\mathscr{C}}_k(X) = \mathscr{L}(X) \setminus \mathscr{C}_k(X)$. We consider that the set $\mathscr{C}_k(X)$ is constructed by arbitrary selection of any k elements from the set $\mathscr{L}(X)$. Theorem 3.2 will derive a bound for the error introduced by subset selection and then we provide a technique to choose the subset in a systematic way.

Theorem 3.2 (Bound for error introduced in compression by subset selection)**.** *Let $p_{\mathscr{L}}(Y \mid X)$ be the density estimated using the machine-derived context set $\mathscr{L}(X)$ for the state X, which is given as*

$$p_{\mathscr{L}}(Y \mid X) = \sum_{i \in \mathscr{L}(X)} \alpha_i K_X(Y, y_i), \tag{9}$$

where $K_X(\cdot,\cdot)$ is a kernel function and α_i is the context prior probability associated with the context $i \in \mathscr{L}(X)$. If $\mathscr{C}_k(X)$ denotes a subset of machine-derived context set $\mathscr{L}(X)$ consisting of k elements, such that $\sum_{i \in \mathscr{C}_k(X)} \alpha_i > 0$, then the density estimate obtained using this subset is given as

$$p_{\mathscr{C}}(Y \mid X) = \sum_{i \in \mathscr{C}_k(X)} \widetilde{\alpha}_i K_X(Y, y_i), \tag{10}$$

where $\widetilde{\alpha}_i = \frac{\alpha_i}{\sum_{i\in\mathscr{C}_k(X)}\alpha_i}$ *is the associated prior. The upper bound of the supremum norm of error in density estimation due to subset selection is proportional to the sum of context priors from the set* $\bar{\mathscr{C}}_k(X)$*, i.e.,* $\mathscr{L}(X)\setminus\mathscr{C}_k(X)$*. In other words,*

$$\|p_{\mathscr{C}}(Y\mid X)-p_{\mathscr{L}}(Y\mid X)\|_\infty \leq \beta_X \sum_{i\in\bar{\mathscr{C}}_k(X)}\alpha_i \tag{11}$$

where $\beta_X\in\mathbb{R}$ *satisfies* $0\leq K_X(\cdot,\cdot)\leq\beta_X<\infty$*.*

Proof. Using (9) and (10), after some algebraic manipulations one can show that the difference in estimates for any $y\in\mathscr{Y}$ is given as

$$p_{\mathscr{C}}(y\mid X)-p_{\mathscr{L}}(y\mid X)=\frac{1}{\sum_{l\in\mathscr{C}_k(X)}\alpha_l}\Bigg(\sum_{i\in\mathscr{C}_k(X)}\sum_{j\in\bar{\mathscr{C}}_k(X)}\alpha_i\alpha_j\big(K_X(y,y_i)-K_X(y,y_j)\big)\Bigg).$$

The supremum norm in this setting of continuous functions is given as

$$\|p_{\mathscr{C}}(Y\mid X)-p_{\mathscr{L}}(Y\mid X)\|_\infty=\sup_{y\in\mathscr{Y}}\big|p_{\mathscr{C}}(y\mid X)-p_{\mathscr{L}}(y\mid X)\big|.$$

Using absolute homogeneity and triangle inequality property for the norm, we obtain

$$\begin{aligned}\|p_{\mathscr{C}}(Y\mid X)-p_{\mathscr{L}}(Y\mid X)\|_\infty &\leq \frac{1}{\sum_{l\in\mathscr{C}_k(X)}\alpha_l}\Bigg(\sum_{\substack{i\in\mathscr{C}_k(X)\\ j\in\bar{\mathscr{C}}_k(X)}}\alpha_i\alpha_j\sup_{y\in\mathscr{Y}}\big|K_X(y,y_i)-K_X(y,y_j)\big|\Bigg)\\ &\leq \frac{1}{\sum_{l\in\mathscr{C}_k(X)}\alpha_l}\Bigg(\sum_{i\in\mathscr{C}_k(X)}\sum_{j\in\bar{\mathscr{C}}_k(X)}\alpha_i\alpha_j\beta_X\Bigg).\end{aligned}$$

Since β_X is the maximum value assumed by the nonnegative-valued kernels. Thus, we obtain the desired result,

$$\|p_{\mathscr{C}}(Y\mid X)-p_{\mathscr{L}}(Y\mid X)\|_\infty\leq\beta_X\sum_{j\in\bar{\mathscr{C}}_k(X)}\alpha_j.$$

□

Remark 3.1 (Optimal k-subset)**.** Since the error upper bound is directly proportional to $(1-\sum_{i\in\mathscr{C}_k(X)}\alpha_i)$, the k-subset with minimum error bound is one for which $\sum_{i\in\mathscr{C}_k(X)}\alpha_i$ is maximum. Thus, if the elements in the context set $\mathscr{L}(X)$ are sorted in descending order of their priors $p(l_i\mid X)$, i.e., α_i, then the best subset $\mathscr{C}_k^*(X)$ corresponds to the first k elements of this sorted sequence. The corresponding error upper bound is $\beta_X(1-\sum_{j\in\mathscr{C}_k^*(X)}\alpha_j)$.

Remark 3.2 (Optimal choice of k)**.** Without loss of generality, we can assume that $\alpha_1\geq\alpha_2\geq\cdots\geq\alpha_{|\mathscr{L}(X)|}$ represents the sorted sequence of context priors. If the error upper bound for a k-subset for any $k\in\{1,2,\ldots,|\mathscr{L}(X)|\}$ is denoted by e_k, then

$e_k = \beta_X(1 - \sum_{j=1}^{k} \alpha_j)$ using result in Remark 3.1. We can verify that, the sequence $\{e_k\}$ is monotonically decreasing with $e_{|\mathscr{L}(X)|} = 0$. These values represent the accuracy of representation of the density. If in certain application we also have a model complexity function $g(k)$, then we can trade-off accuracy with complexity using a criterion, such as Akaike Information Criterion [12], to find the optimal value of k, which minimizes the chosen criterion.

The error bound derived in Theorem 3.2 is usually conservative, but this conservative analysis leads to a simple expression for e_k, which can readily be evaluated for all $k \in \{1, 2, \ldots, |\mathscr{L}(X)|\}$. Unlike the technique in Section 3.1, the subset selection approach can give a relationship of subset size or compression ratio with maximum error in estimation, which in turn can lead to choosing appropriate compression as shown in Remark 3.2. However, the subset selection approach directly ignores the contexts with low priors and does not use information of overlap/distance between individual components, which might not be desirable for certain applications.

This section presented two techniques for compression of context sets along with the main results of the upper bound of error due to approximation. The error bound evaluated in first approach was in terms of statistical distance functions, whereas in second case we derive a more intuitive bound in terms of the supremum norm. In the next section, we will use these techniques for cardinality reduction of context sets derived from multiple seismic sensor data for a target classification problem.

4 Experiments and Results

The procedure and results of experimental validation of the context set compression techniques are presented in this section. We conducted field experiments to collect data from unattended ground sensors, such as seismic, acoustic, and passive infrared sensors, for a border-crossing target detection and classification problem. In this study, we use time-series data from two different seismic sensors which were separated by 7 meters and the target is passing almost parallel to the line joining the two sensors at various distances between 2 to 8 meters. The hypothesis set consists of human target walking ($x = 1$) and human target running ($x = 2$) class and the goal is to classify the activity of the target using data from both the seismic sensors, as shown in Figure 2.

The dataset consists of 110 runs for walking and 118 runs for running. We partition the sample into a training set and testing set consisting of 60% and 40% of the data respectively. All results are generated for 10 different partitions of the sample and average results are available for the different steps. In the first step, low-dimensional features are extracted from time-series data using *symbolic dynamic filtering* (SDF) [19]. In SDF, we partition the measurement space into several regions and assign a symbol to each region. The set of these symbols is known as the alphabet. Bias is removed from the measurement time-series data to make it zero-mean and it is also normalized to have unit variance to remove the effect of target distance on signal amplitude. The resulting time-series data is then represented by a

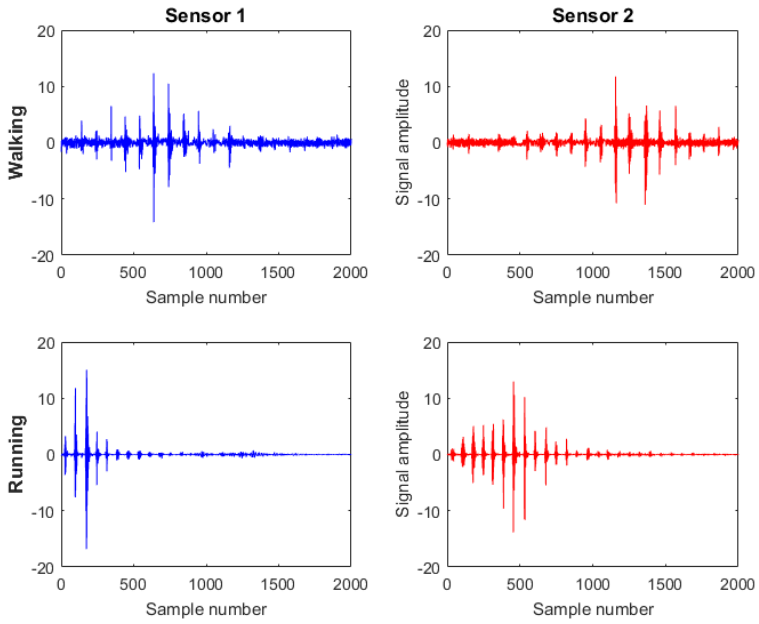
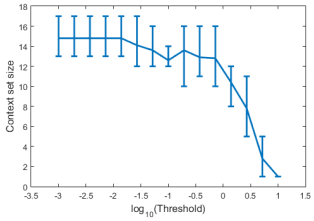


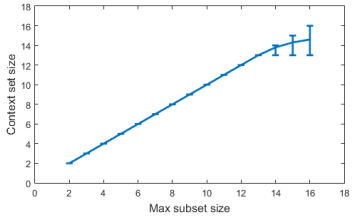
Fig. 2: Seismic sensor time series data for walking and running classes

symbol sequence and the statistics of evolution of this sequence is represented by a D-Markov model [20]. In this analysis, we used alphabet size of 6 and depth D of 2, resulting in a D-Markov model of 7 states after state-splitting and state-merging. The left eigenvector of state transition matrix of the D-Markov model corresponding to the eigenvalue of 1 is the stationary state probability vector, which is used as a low-dimensional feature vector for each time-series data. The details on the D-Markov model construction and feature extraction techniques are given in [20].

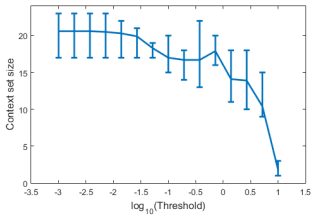
The second step is of *unsupervised context learning* which uses nonparametric density estimation for obtaining machine-derived context sets from kernel-based mixture models as shown in Section 2. The density estimation process is used for computing the joint likelihood of obtaining a feature Y_1 from seismic sensor 1 and a feature Y_2 from seismic sensor 2, given that the state is X. The kernels used in the mixture modeling process are Gaussian with diagonal covariance matrix having identical entries, i.e., $K_i(y,y_i) = (2\pi\gamma)^{-d_i/2}\exp(-\frac{(y-y_i)^{\mathrm{T}}(y-y_i)}{2\gamma^2})$, where d_i is the dimensionality of feature Y_i for $i = 1,2$ and γ is the kernel shape parameter. Using $\gamma = 0.01$, the resulting context sets for state 1 have cardinality (i.e., number of elements in the set) has mean 14.80 and standard deviation 1.47, whereas for state 2, the cardinality has mean 20.60 and standard deviation 1.65. This analysis uses maximum likelihood decision rule for classification that gives state estimate as



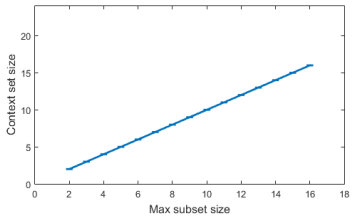
(a) MCE for $x = 1$



(b) Subset selection for $x = 1$

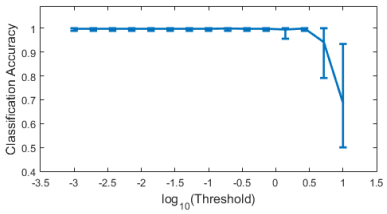


(c) MCE for $x = 2$

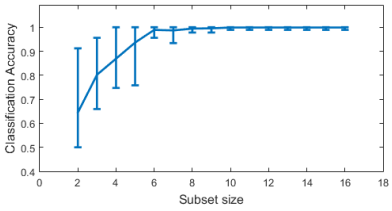


(d) Subset selection for $x = 2$

Fig. 3: Mean and range of cardinality of the compressed context set



(a) MCE



(b) Subset selection

Fig. 4: Mean and range of classification accuracy with compressed context sets

$$\hat{x} = \arg\max_{x \in \mathscr{X}} p(Y_1, Y_2 \mid x) = \arg\max_{x \in \mathscr{X}} \sum_{c \in \mathscr{C}(x)} p(c|x) p(Y_1, Y_2 \mid x, c). \tag{12}$$

For the machine-derived context set with $\gamma = 0.01$, the classification accuracy was 99.78% on average using the decision rule in (12).

In the third step, we use the two proposed *context set cardinality reduction* techniques to obtain the compressed context sets. In the maximal clique enumeration (MCE) technique, the contextual observation densities are multivariate Gaussian distributions with mean $\mu_i(x)$ and identical covariance matrix $\Sigma_\gamma(x)$, which is parametrized by the kernel shape parameter γ, thus $p(Y \mid X = x, L = l_i) \sim \mathscr{N}(\mu_i(x), \Sigma_\gamma(x))$. In order to construct the weight matrix, we use the closed form expression of the Bhattacharyya distance for Gaussian densities [16], given as

$$w_{ij}(x) = d\big(p(Y \mid X = x, L = l_i), p(Y \mid X = x, L = l_j)\big)$$
$$= \frac{1}{8}(\mu_i(x) - \mu_j(x))^{\mathrm{T}} \Sigma_\gamma(x)^{-1} (\mu_i(x) - \mu_j(x)) \tag{13}$$

for $i, j = 1, 2, \ldots, |\mathscr{L}(x)|$. The threshold parameter ε to be used in the `ConstructGraph` function of the MCE approach is varied from 10^{-3} to 10^{1} in 15 equal steps in the log scale. For the graph obtained from the `ConstructGraph` function, we perform the maximal clique enumeration process and compute minterms of the obtained set of cliques. Note that as threshold increases, the cardinality of the compressed set shows a non-monotonic reducing trend in Figure 3a and 3c as number of cliques need not reduce monotonically with reduction of the edge set of the graph. The `Minterms` procedure ensures by definition that the number of cliques in the resulting set is upper-bounded by cardinality of the machine defined context set, that is, $|\mathscr{C}(x)| \leq |\mathscr{L}(x)|$ for all $x \in \mathscr{X}$, thus, we will get some compression. The classification performance summary using compressed context sets using MCE is given in Figure 4a. The results show that for $\varepsilon = 10^{0.42} = 2.68$, the mean cardinality of context sets is $|\mathscr{C}(1)| = 7.8$ and $|\mathscr{C}(2)| = 13.9$, and the average classification accuracy is same as the full context set. This result demonstrates that cardinality reduction need not significantly affect the class performance. However, reducing cardinality further by increasing ε leads to significant deterioration in performance in this case. Cross-validation can be used to choose the appropriate value for the threshold ε.

In the subset selection approach, the maximum size of the subset, denoted by k, is varied from 2 to 16. If the original context set is smaller than the chosen set size, we do not perform any other computation, else we choose the best k-subset using Remark 3.1. Thus, Figure 3b and 3d shows a monotonic trend of context set size with the chosen parameter k. The classification performance shows an increasing trend with the size of context set. For $k = 8$, the performance is 99.56% and for $k > 8$, the performance is as good as the original set. Thus, compression of context sets can be achieved by subset selection techniques as well. A suitable context set size can be chosen by using cross-validation, if classification accuracy is the selection criterion, else one can use the method outlined in Remark 3.2 to choose context set size.

5 Conclusion

This chapter presents two different approaches to control the size of context sets in an unsupervised learning setting. Learning approaches with density estimation to obtain machine-defined context set from multi-modal sensor data is reviewed in this chapter and the resulting density estimate is used in both the proposed approaches. One approach relies on the graph-theoretic concept of maximal clique enumeration to identify contexts which affect the sensor data in a similar way and it creates approximate equivalence classes of the machine-defined contexts. The upper bound of error introduced by this compression was identified. A subset selection

approach is also presented in this chapter and the upper bound of error introduced by subset selection is derived. In this approach, the prior probabilities over context played an important role to obtain the best subset. We could derive a conservative relation between the error upper bound and cardinality of the context set. These approaches were then used with seismic sensor data collected in field experiments for walking-type classification of border crossing targets. The results validate that these techniques are indeed useful for compression of context sets and one can maintain similar classification performance with a much smaller context set. In future, an agglomerative clustering approach that can provide an estimate of the error introduced by compression will be explored to find a computationally inexpensive approach to allow representation of data from all relevant regions in the measurement space.

References

1. F. Darema, "Dynamic data driven applications systems: New capabilities for application simulations and measurements," in *Computational Science–ICCS 2005*, pp. 610–615, Springer, 2005.
2. B. Kahler, E. Blasch, and L. Goodwon, "Operating condition modeling for ATR fusion assessment," in *Defense and Security Symposium*, pp. 65710D–65710D, International Society for Optics and Photonics, 2007.
3. N. Virani, J.-W. Lee, S. Phoha, and A. Ray, "Learning context-aware measurement models," in *American Control Conference (ACC), 2015*, pp. 4491–4496, IEEE, 2015.
4. C. R. Ratto, P. Torrione, and L. M. Collins, "Exploiting ground-penetrating radar phenomenology in a context-dependent framework for landmine detection and discrimination," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 5, pp. 1689–1700, 2011.
5. S. Phoha, N. Virani, P. Chattopadhyay, S. Sarkar, B. Smith, and A. Ray, "Context-aware dynamic data-driven pattern classification," *Procedia Computer Science*, vol. 29, pp. 1324–1333, 2014.
6. E. Blasch, J. Nagy, A. Aved, E. K. Jones, W. M. Pottenger, A. Basharat, A. Hoogs, M. Schneider, R. Hammoud, G. Chen, *et al.*, "Context aided video-to-text information fusion," in *Information Fusion (FUSION), 2014 17th International Conference on*, pp. 1–8, IEEE, 2014.
7. L. Snidaro, J. García, J. Llinas, and E. Blasch, *Context-Enhanced Information Fusion: Boosting Real-World Performance with Domain Knowledge*. Springer, 2016.
8. N. Virani, J.-W. Lee, S. Phoha, and A. Ray, "Dynamic context-aware sensor selection for sequential hypothesis testing," in *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pp. 6889–6894, Dec 2014.
9. E.Blasch, J.Herrero, L.Snidaro, J.Llinas, G.Seetharaman, and K.Palaniappan, "Overview of contextual tracking approaches in information fusion," in *Proceedings of SPIE*, 2013.
10. S. Mukherjee and V. Vapnik, "Support vector method for multivariate density estimation," *Center for Biological and Computational Learning. Department of Brain and Cognitive Sciences, MIT. CBCL*, vol. 170, 1999.
11. N. Virani, J.-W. Lee, S. Phoha, and A. Ray, "Information-space partitioning and symbolization of multi-dimensional time-series data using density estimation," in *American Control Conference (ACC), 2016*, pp. 3328–3333, IEEE, 2016.
12. C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.
13. C. Bron and J. Kerbosch, "Algorithm 457: finding all cliques of an undirected graph," *Communcations of ACM*, vol. 16, no. 9, pp. 575–577, 1973.
14. E. Tomita, A. Tanaka, and H. Takahashi, "The worst-case time complexity for generating all maximal cliques and computational experiments," *Theoretical Computer Science*, vol. 363, no. 1, p. 2842, 2006.

15. J. R. Hershey and P. Olsen, "Approximating the Kullback Leibler divergence between gaussian mixture models," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–317, IEEE, 2007.
16. F. J. Aherne, N. A. Thacker, and P. I. Rockett, "The Bhattacharyya metric as an absolute similarity measure for frequency coded data," *Kybernetika*, vol. 34, no. 4, pp. 363–368, 1998.
17. J. W. Moon and L. Moser, "On cliques in graphs," *Israel Journal of Mathematics*, vol. 3, no. 1, pp. 23–28, 1965.
18. D. Avis and K. Fukuda, "Reverse search for enumeration," *Discrete Applied Mathematics*, vol. 65, no. 1–3, pp. 21–46, 1996.
19. A. Ray, "Symbolic dynamic analysis of complex systems for anomaly detection," *Signal Processing*, vol. 84, pp. 1115–1130, July 2004.
20. K. Mukherjee and A. Ray, "State splitting and merging in probabilistic finite state automata for signal representation and analysis," *Signal processing*, vol. 104, pp. 105–119, 2014.