Exploiting Environmental Information for Improved Underwater Target Classification in Sonar Imagery

David P. Williams and Elias Fakiris

***Abstract*—In many remote-sensing applications, measured data are a strong function of the environment in which they are collected. This paper introduces a new context-dependent classification algorithm to address and exploit this phenomenon. Within the proposed framework, an ensemble of classifiers is constructed, each associated with a particular environment. The key to the method is that the relative importance of each object (i.e., data point) during the learning phase for a given classifier is controlled via a modulating factor based on the similarity of auxiliary environment features. Importantly, the number of classifiers to learn and all other associated model parameters are inferred automatically from the training data. The promise of the proposed method is demonstrated on classification tasks seeking to distinguish underwater targets from clutter in synthetic aperture sonar imagery. The measured data were collected with an autonomous underwater vehicle during several large experiments, conducted at sea between 2008 and 2012, in different geographical locations with diverse environmental conditions. For these data, the environment was quantified by features (extracted from the imagery directly) measuring the anisotropy and the complexity of the seabed. Experimental results suggest that the classification performance of the proposed approach compares favorably to conventional classification algorithms as well as state-of-the-art context-dependent methods. Results also reveal the object features that are salient for performing target classification in different underwater environments.**

***Index Terms*—Autonomous underwater vehicles (AUVs), context-dependent classification, environmental dependence, mine countermeasures, synthetic aperture sonar (SAS).**

I. Introduction

AN IMPLICIT assumption made in most statistical learning algorithms [1] is that training data and test data are generated from the same underlying distribution. That is, the labeled data used to train a classifier will be representative of the unlabeled test data for which predictions must subsequently be made. Therefore, if data mismatch does exist between these two sets, classification performance will suffer accordingly. In many remote-sensing applications, the assumption of data homogeneity is violated because of a strong dependence on the environment in which the data are collected.

Manuscript received October 9, 2013; revised November 7, 2013; accepted December 17, 2013. Date of publication January 10, 2014; date of current version May 22, 2014.

D. P. Williams is with the Centre for Maritime Research and Experimentation, NATO Science and Technology Organization, 19126 La Spezia, Italy (e-mail: williams@cmre.nato.int).

E. Fakiris is with the Laboratory of Marine Geology and Physical Oceanography, University of Patras, 26504 Rio, Greece (e-mail: fakiris@upatras.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2013.2295843

The underwater target classification task concerned with discriminating mines from clutter in sonar imagery is known to suffer from this phenomenon. For example, it is common to encounter different types of clutter objects (e.g., rocks, vegetation, and man-made junk) at different geographical locations. However, differing feature distributions can also be attributed to a more subtle cause: seabed composition. Consider a feature that measures the height of an object on the seafloor based on the length of the shadow cast (and the range from, and altitude of, the imaging sonar). This feature may be expected to be relatively robust as it is tied to a physical property of the object. However, in seabeds that are composed of soft clay, it is possible that objects will sink into the seabed and become partially buried, thereby decreasing the observable object height. In contrast, on a seabed of hard-packed sand, objects are likely to be proud on the seafloor, so the measured heights will be correspondingly larger. The result is that the height-feature measurement *for the same given object* would be very different in these two environments. More generally, the values of template-based features [2], [3], segmentation-based features [4], [5], and other commonly used object features [6] for which the calculations intrinsically depend on the surrounding seabed can also be strongly affected by environmental conditions.

Substantial research has explored various versions of the transfer learning problem, which seeks to improve classification performance when the underlying distributions generating training data and (future) test data differ, but most of these approaches require access to the test data [7]–[10]. Instead, our scenario is more similar to the one addressed by mixture of experts [11] algorithms, which learn multiple classifiers (i.e., "experts") that are each trusted to different degrees in different regions of feature space. However, the approaches developed in this vein [12], [13] cannot address the case described previously, in which unique classifiers are required not in different regions of feature space but rather in different *environments* (or regions of what can be called "meta-feature space"). To our knowledge, only one strand of work, initiated in [14], has addressed our specific setting of context-dependent classification.

In [14], it was assumed that context labels denoting the environment were possessed, which enabled a supervised approach for learning models for auxiliary context features. However, acquiring such labels in remote-sensing applications can be difficult or even impossible, so the framework was extended in [15] and [16] to handle the case in which context labels are not available. This is also the scenario that we address in this work, since obtaining context labels in the underwater target classification problem is impractical. (Assigning context labels derived from geographical or data-collection site is also

not feasible because the intrasite variation of the underwater environment is too great.)

In the framework of [15] and [16] and in our proposed approach, the class prediction for a data point is made as a weighted sum of an ensemble of classifier predictions, with the weighting dictated by the object's similarity to each environment. The differences between the approaches lie in how the classifiers are learned and how the probability of being associated with each environment is defined. A nonparametric Bayesian approach is employed in [15] and [16], whereas our independently formulated algorithm is inspired by maximum entropy methods. Another difference between the previous research and our work lies in the application domain addressed: Instead of performing landmine detection with ground-penetrating radar or hyperspectral imagery [15], [16], we tackle the task of classifying underwater targets with sonar data.

The present work introduces a new classification framework that compensates for data mismatch by first quantifying the environmental conditions under which each data point is collected. This auxiliary information is then incorporated into a learning process that constructs an ensemble of classifiers. The key is that the relative importance of each object (i.e., data point) during the learning phase for a given classifier is controlled via a modulating factor computed by comparing the object's environment features with analogous environment features assigned to each classifier. To enhance the rigor of this weighting, we appeal to the idea of Boltzmann distributions and the concept of energy states of a system. In this analogy, the probability that the system is in a specified state is equivalent to the contribution of an object to the learning process of the specified classifier. Just as low-energy states of a system are more probable, the contribution of an object to a classifier will be stronger when the environments associated with the object and classifier have low dissimilarity. All free parameters, including the number of classifiers to learn, are inferred automatically by appealing to maximum entropy methods, for which strong connections to information theory exist [17]. Employing the auxiliary environment meta-features enables a level of flexibility not possible in traditional classification algorithms, namely, that the prediction at a particular data point in feature space can be different for different environments.

The remainder of this paper is organized as follows. The proposed classification algorithm that exploits auxiliary environmental information is described in Section II. The measured sonar data sets used in the subsequent study are summarized in Section III, with experimental results presented in Section IV. Concluding remarks and directions for future work are noted in Section V.

II. Proposed Classification Algorithm

A. Algorithm Derivation

The proposed classification algorithm exploiting auxiliary environmental information is outlined in detail here. To avoid interrupting the flow of the derivation, a more thorough discussion explaining the rationale surrounding various aspects is withheld until Section II-B. For the sake of clarity, we first present the algorithm assuming that the environment is represented by a single scalar meta-feature; later, we present the extension to the general case of a vector of meta-features.

1) Preliminaries: Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote a (column) vector of d features representing the ith object of a training set of N such objects. Let $z_i \in \mathbb{R}$ denote a scalar meta-feature that quantifies auxiliary information about the conditions under which the ith object was collected. We refer to this meta-feature as the *environment* feature. Let $y_i \in \{1, -1\}$ denote the class label that corresponds to the ith object $\boldsymbol{x}_i$. Collect the N sets of object features, class labels, and environment features as $\mathbf{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, $Y = \{y_i\}_{i=1}^N$, and $Z = \{z_i\}_{i=1}^N$, respectively.

The objective is to perform binary classification using the training data $\{\mathbf{X}, Y, Z\}$, where the presence of the auxiliary environment information Z distinguishes the task from standard supervised classification tasks. It should be noted that the proposed algorithm is a purely supervised approach, assuming no knowledge of, or access to, the test data on which classification is to be performed subsequently.

2) Establishing Data Importance: In the proposed algorithm, rather than learning a single classifier, a set of $C > 2$ classifiers is learned. The jth such classifier will be associated with an assigned environment feature value z_j^*. The specification of C and the construction of the set $Z^* = \{z_j^*\}_{j=1}^C$ will be addressed in the following discussion.

A weight modulating the relative importance of the ith object during the learning of the jth classifier is calculated using the Boltzmann distribution

$$\omega(z_i, z_j^*) = \frac{\exp\left\{-d(z_i, z_j^*)/\beta\right\}}{\sum_{k=1}^{C} \exp\left\{-d(z_i, z_k^*)/\beta\right\}} \tag{1}$$

where $\beta > 0$ is a fixed scaling parameter and

$$d\left(z_i, z_j^*\right) = \left|z_i - z_j^*\right| \tag{2}$$

is the distance between the ith object's environment feature and the environment feature associated with the jth classifier. It is here that the key auxiliary environment information is exploited. It should also be noted that the denominator in (1) is a normalizing constant ensuring $0 < \omega(z_i, z_j^*) < 1$.

3) Parameter Selection: The aforementioned formulation relies on three as-yet-unspecified quantities: C, the number of classifiers to be learned; Z^*, the set containing the environment feature associated with each classifier; and the scaling parameter β. These quantities are determined in the following manner. The first and last elements of Z^* are set to be the smallest and largest values of the training set's environment features, respectively

$$z_1^* = \min_i z_i \tag{3}$$

$$z_C^* = \max_i z_i. \tag{4}$$

The remaining $C - 2$ elements z_j^* are then assigned values that divide $[z_1^*, z_C^*]$ into equal partitions. By selecting the extrema of Z for inclusion in Z^*, the learned classifiers will span the greatest range of potential test data environment values (which are unknown *a priori*) for which training data exists.

Next, $\beta \in \mathbb{R}^+$ and $C \in \mathbb{Z}^+$ are determined jointly by performing a brute-force (yet very easy and fast) search to find the (β, C) pair that maximizes the entropy of the importance

weights $\omega(z_i, z_j^*)$, calculated using all N training data points. (Recall that $\omega(z_i, z_j^*)$ depends on both β and C.) That is, for a given (β, C) pair, the entropy

$$H(\omega|\beta, C) = -\sum_{\omega_k \in \Omega_\Delta} p(\omega_k) \log_2 p(\omega_k) \tag{5}$$

is calculated, where $p(\omega_k)$ is the relative frequency with which the (continuous-valued) weights $\omega(z_i, z_j^*)\ \forall\, i, j$ are mapped to the kth element in the discrete alphabet of quantized weights Ω_Δ. The (β, C) pair that maximizes the entropy of the weights (5) is then selected.

4) Learning of Classifiers: With C selected, all z_j^* are specified, and therefore, all $d(z_i, z_j^*)$ can be computed via (2). With β determined, all $\omega(z_i, z_j^*)$ can be readily computed as well via (1). Let $\Omega_{(j)} = \{\omega(z_i, z_j^*)\}_{i=1}^N$ be the set of weights associated with the training data $\{\mathbf{X}, Y, Z\}$ for the jth classifier. The jth classifier is then learned using $\{\mathbf{X}, Y, \Omega_{(j)}\}$—the information contained in Z having been fully transferred to $\Omega_{(j)}$—by modulating the contribution of the ith object $\boldsymbol{x}_i$ by $\omega(z_i, z_j^*)$ in the (base) classifier's objective function. This weighting effectively controls the trust placed in each data point for the given classifier.

Many standard classification algorithms can be employed here as the base classifier within this framework, but we do assume that the classifier used will produce probabilistic predictions. In the experiments presented here, we use a modified form of the relevance vector machine (RVM) [18] with no kernel,[1] which coincidentally the state-of-the-art context-dependent algorithms [14]–[16] also employ, thereby making direct performance comparisons feasible.

With this kernel choice, the classifier parameters are weights on the features themselves rather than on basis functions, which means the learned parameters can be analyzed in terms of feature selection, the implications of which will be discussed later. The RVM is also convenient because it provides probabilistic predictions that can be easily combined. Moreover, employing the RVM requires only a minor modification to the original objective function and, for the learning phase, its gradient and Hessian with respect to the classifier parameters $\boldsymbol{w}_{(j)}$. (This particular modification is straightforward and does not affect the theoretical properties of the RVM, but care must be taken to ensure the same if one chooses to use a different base classification method.) The modified RVM objective function to be maximized under the proposed framework for the jth classifier becomes

$$J_{(j)} = \sum_{i=1}^{N} \omega(z_i, z_j^*) \log \sigma\left(y_i \boldsymbol{w}_{(j)}^\top \boldsymbol{x}_i\right) - \frac{1}{2} \boldsymbol{w}_{(j)}^\top \mathbf{A}_{(j)} \boldsymbol{w}_{(j)} \tag{6}$$

where $\mathbf{A}_{(j)}$ is a diagonal matrix of hyperparameters associated with the sparsity-promoting prior within the RVM model, $\boldsymbol{w}_{(j)}$ is the vector of classifier parameters to be learned, and $\sigma(u) = (1 + \exp\{-u\})^{-1}$ is the sigmoid function. (To recover the original objective function, one must simply remove the $\omega(z_i, z_j^*)$ factor.) Standard classifier learning is then undertaken as one normally would; the culmination of this process for the jth classifier is the vector of learned classifier parameters $\boldsymbol{w}_{(j)}$.

5) Prediction: Let $\mathbf{W} = \{\boldsymbol{w}_{(j)}\}_{j=1}^C$ collect all of the learned classifiers. Then, given a new unlabeled test object $\boldsymbol{x}_\ell$ with environment meta-feature z_ℓ, class prediction is made using a weighted average of the C classifiers' predictions; this weighting is again specified by $\omega(z_\ell, z_j^*)$, measuring the similarity of the test object's environment feature with each classifier's environment feature, computed using (1). Thus, the probability that test object $\boldsymbol{x}_\ell$ belongs to class $y_\ell = 1$ is given by

$$p(y_\ell = 1|\boldsymbol{x}_\ell, z_\ell, \mathbf{W}) = \sum_{j=1}^{C} \omega(z_\ell, z_j^*) p\left(y_\ell = 1|\boldsymbol{x}_\ell, \boldsymbol{w}_{(j)}\right) \tag{7}$$

where $p(y_\ell = 1|\boldsymbol{x}_\ell, \boldsymbol{w}_{(j)})$ is the prediction of the jth classifier.

For the modified RVM used in this work, $p(y_\ell = 1|\boldsymbol{x}_\ell, \boldsymbol{w}_{(j)}) = \sigma(\boldsymbol{w}_{(j)}^\top \boldsymbol{x}_\ell)$.

6) Extension: Multiple Environment Meta-Features: The aforementioned algorithm can easily be extended to handle the case in which the environment is represented by multiple meta-features, rather than a single scalar meta-feature. Let $\boldsymbol{z}_i = [z_{i1}\ \ z_{i2}\ \ \cdots\ \ z_{iF}]^\top$ denote a vector of F meta-features that quantifies auxiliary information about the conditions under which the ith object was collected.

The weight modulating the relative importance of the ith object during the learning of the jth of C total classifiers is then modified to be calculated as

$$\omega(\boldsymbol{z}_i, \boldsymbol{z}_j^*) = \frac{\exp\left\{-\sum_{f=1}^F d(z_{if}, z_{jf}^*)/\beta_f\right\}}{\sum_{k=1}^C \exp\left\{-\sum_{f=1}^F d(z_{if}, z_{kf}^*)/\beta_f\right\}} \tag{8}$$

where $\beta_f > 0$ is a fixed scaling parameter and

$$d\left(z_{if}, z_{jf}^*\right) = \left|z_{if} - z_{jf}^*\right| \tag{9}$$

is the distance between the ith object's fth environment feature and the fth environment feature associated with the jth classifier. The distance calculation is made feature-by-feature, and a unique scaling parameter is included for each meta-feature to prevent the contribution of one feature from unfairly dominating.

In the case of a scalar meta-feature, C was the number of classifiers to be learned because there were C unique meta-feature values associated with the classifiers. When there is a vector of meta-features, C_f will correspond to the number of unique classifier-associated values of the fth meta-feature. The assumption is that the vector of meta-feature values associated with a given classifier will be formed from the Cartesian product of the individual meta-feature value sets. This means that the total number of classifiers to be learned will be $C = \prod_{f=1}^F C_f$. Despite the unfavorable scaling, if the meta-feature dimension F is low, jointly learning the unknowns C_f and β_f for all f by maximizing the entropy of the weights will still be feasible (the brute-force search used in the scalar meta-feature case is also used here). If F is large, dimensionality reduction techniques such as principal component analysis [19] or locally linear embedding [20] can be employed to first map the meta-feature data into a lower-dimensional space. (In fact,

[1] That is, the RVM kernel function intended to measure the similarity between two data points is instead defined as $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{x}_i$, which has been referred to as a "direct kernel" [14].

a similar approach was employed in [14] to reduce the meta-feature dimension, via linear discriminant analysis.)

Once all C_f, β_f, and $z^*_{\cdot f}$ are obtained, classifier learning and prediction proceeds as in the scalar meta-feature case.

B. Discussion

The principal insight being leveraged in the proposed framework is that the values of features extracted to represent objects at a given site can be strongly influenced by (and correlated with) a meta-feature summarizing environmental properties of the area. This environmental dependence is exploited by learning an ensemble of classifiers, each associated with a particular environment. The data that are used to learn each classifier are automatically weighted according to their relevance (i.e., similarity) to the environment under consideration. In this way, all available data are always used to learn each classifier, yet classifier diversity (across different environments) is still achievable via unique weightings.

It should be noted that the normalization in (1) ensures that the total (summed) weight associated with each data point is unity. That is, although a given data point may contribute a different amount to each classifier, each data point will, in aggregate, contribute the same amount to the overall training process. Therefore, to paraphrase George Orwell, all data are equal, but some data are more equal than others.

The measure of dissimilarity d in (2) is equivalent to the "state energy" in the Boltzmann distribution. Very simply, when a data point's environment feature is similar to the environment feature value associated with a given classifier—i.e., in low-energy states—that data point will be trusted more in that classifier's learning process. When a data point's environment feature is very dissimilar—i.e., in high-energy states—the relative importance of that data point in the learning process will be small. In this sense, one can view d as quantifying the energy needed to generate a given data point in a particular environment.

To determine the values of the scaling parameter β and the number of classifiers C, the entropy of the weights ω was maximized. Maximum entropy methods have a strong theoretical underpinning from information theory [17], where they are shown to assume the least about unknowns. However, there is also an intuitive rationale for the decision to exploit them in this context. When the entropy of the weights is maximized, the diversity of the different learned classifiers will tend to be large because the contributions (weights) associated with each data point will be highly varied. This classifier diversity is important because we want to encourage different classifiers to be learned in different environments (as much as the data can support such a result). If β is too small, each data point will have one weight near unity and all others near zero. In this case, effectively, each classifier would be learned using only a subset of the data (namely, the data points whose environment is most similar to the classifier's under consideration). If β is too large, each data point will have virtually equal contributions (weights) for each classifier. As a result, each classifier learned would be nearly identical, thereby eliminating any potential for improved performance.

Because the environmental meta-feature is a continuous variable, we discretize this space into C values. The discretization is particularly important because, if C is too small, the classifiers will not be tailored finely enough to the environment of interest. Similarly, if C is too large, the contribution of each data point to learning each classifier will be weakened, in turn decreasing the data set diversity among the classifiers, and the resulting classifiers will be too similar. In the Boltzmann distribution analogy referenced earlier, C, the number of different environments possible, is the number of possible states of the system.

One of the particularly attractive aspects of the proposed algorithm is that there are no free parameters ("knobs") that must be tuned or tweaked. All of the necessary quantities are automatically learned from the data by the algorithm itself.

C. Illustrative Example: Synthetic Data

To help illustrate the power of the proposed algorithm, we first consider binary classification of a synthetic 2-D data set, with the low feature dimension purposely chosen to permit easy visualization. Let $y_i \in \{1, 0\}$ be the class label of a data point represented by the feature vector $\boldsymbol{x}_i \in \mathbb{R}^2$ that is associated with the environment meta-feature $z_i \in \mathbb{R}^1$.

We consider the case in which the distribution of the feature data is a strong function of the meta-feature. Specifically, the data are generated according to

$$y_i \sim \mathcal{B}(0.5) \tag{10}$$

$$z_i \sim \mathcal{U}(0, 1) \tag{11}$$

$$\boldsymbol{x}_i | z_i, y_i \sim \mathcal{N}(\mu_{y_i}, 0.5\mathbf{I}) \tag{12}$$

where $\boldsymbol{\mu}_1 = [z_i \quad z_i]^\top$, $\boldsymbol{\mu}_0 = [1 - z_i \quad 1 - 2z_i]^\top$, and $\mathbf{I}$ is the 2-D identity matrix; $\mathcal{B}$, $\mathcal{U}$, and $\mathcal{N}$ indicate Bernoulli, uniform, and normal distributions, respectively, for which the parameter notation should be obvious. It can be observed that the meta-feature essentially induces a migration in feature space of the data from each class.

A data set of 1000 points is generated and then randomly divided into two disjoint sets of equal size, one treated as labeled training data and the other treated as unlabeled test data. The complete data set is shown in Fig. 1, where it can be seen that the two classes cannot be discriminated easily. Moreover, augmenting the feature space by treating z_i as a third feature does not alleviate the challenge. However, by *conditioning* on z_i, discrimination becomes more feasible.

We compare the classification performance of three approaches: 1) a standard classifier constructed on the 2-D feature data; 2) a classifier constructed on the augmented 3-D feature space (FS) that includes the meta-feature z_i as a third feature; 3) the proposed method. A modified RVM with a "direct kernel" is employed as the base classifier for all methods in these experiments.

As explained in Section II-A3, the proposed method sets C and the scaling parameter β by maximizing the entropy of the weights. For this data set, the learned values of the parameters were $C = 3$ and $\beta = 0.3$. The decision boundaries for the three classifiers are shown in Fig. 2, where it can be observed that the classifiers associated with the different meta-feature values are quite distinct.

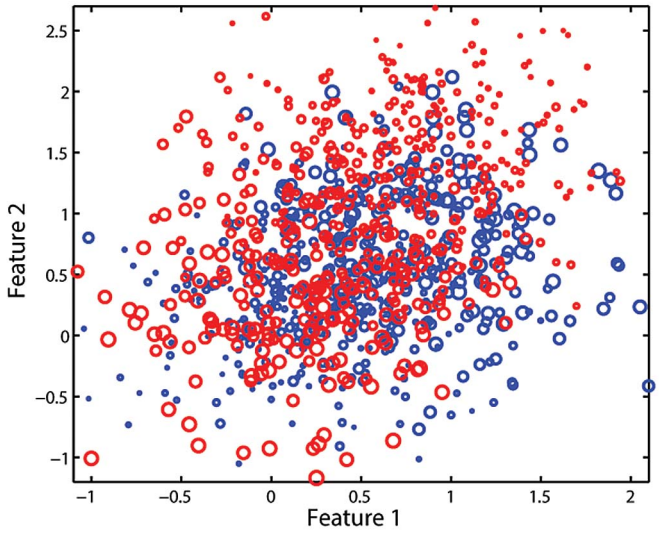

Fig. 1. Set of 2-D synthetic data (with class 1 represented by blue circles and class 0 represented by red circles), with the value of the environmental meta-feature proportional to the circle size.

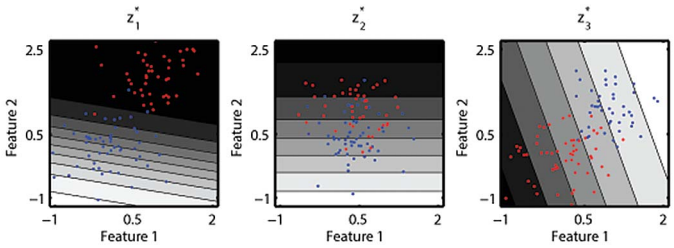

Fig. 2. Proposed method's $C = 3$ learned classifiers for the synthetic data. The background shows the classifier prediction of belonging to class 1 (higher probability corresponds to lighter shades) everywhere in feature space, assuming that the test data point's meta-feature is equal to the given z_j^*. For context, the 100 data points whose environment meta-feature is most similar to each classifier's environment meta-feature are also shown.

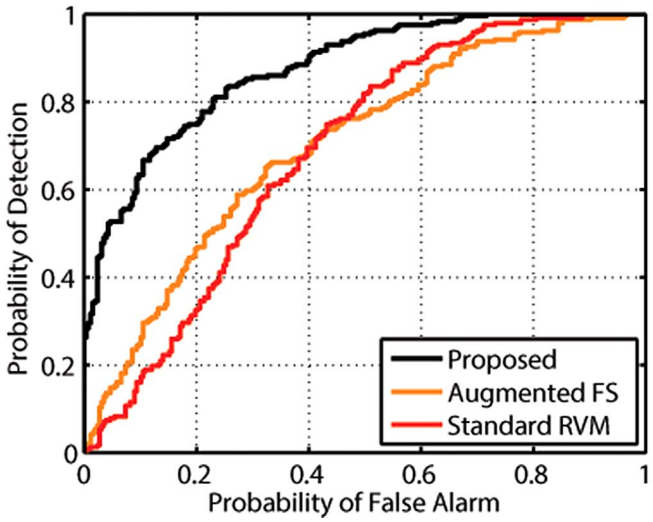

Fig. 3. Classification performance on the synthetic data set.

The performance of the three methods considered is shown in the form of receiver operating characteristic (ROC) curves in Fig. 3.

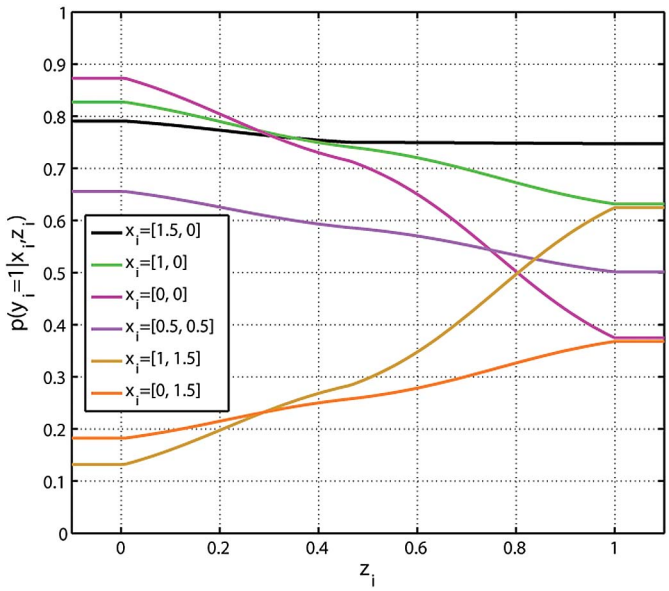

Fig. 4. For six specific data points in (object) feature space, the predicted probability of belonging to class 1 shown as a function of possible environment meta-feature values.

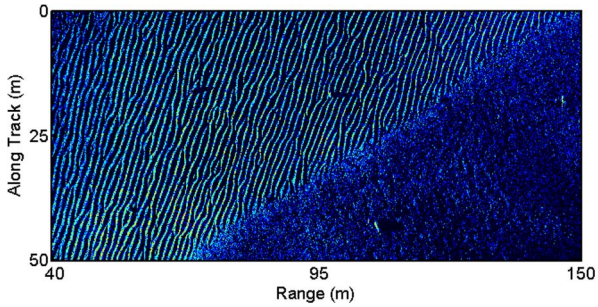

Fig. 5. Example SAS image, collected by the MUSCLE AUV, that contains seven targets, some of which are located in sand ripple fields.

As can be seen from the figure, the proposed method performs much better than the competing methods. The approach using an augmented feature space (treating the environment meta-feature as a third object feature) still fails because no single classifier can reliably discriminate the classes. In contrast, the proposed method exploits the auxiliary environment information in a fundamentally different manner. The result is that the data can be classified much more accurately because the algorithm has the flexibility to make different predictions at the same point in feature space for different environment values. This is highlighted in Fig. 4, which shows how the proposed method's predictions vary as a function of the meta-feature value z_i at specific locations of $\boldsymbol{x}_i$.

III. Sonar Data Sets

Synthetic aperture sonar (SAS) works by coherently summing received acoustic signals of overlapping elements in an array, and it provides an order-of-magnitude improvement in resolution over simple (real aperture) side-scan sonar data [21]. The resulting high-resolution SAS imagery provides a detailed view of the seafloor that makes detection of proud targets possible. Typically, a target in SAS imagery will exhibit a

TABLE I
DETAILS OF SEA EXPERIMENTS DURING WHICH SONAR DATA WERE COLLECTED

SEA EXPERIMENT - AREA	LOCATION	YEAR	DAYS/MONTH	NOTABLE ENVIRONMENTAL CHARACTERISTICS
COLOSSUS 2 - D	RIGA, LATVIA	2008	15-19/4, 3/5	CLAY, FLAT SAND
COLOSSUS 2 - B	LIEPAJA, LATVIA	2008	26/4-2/5	FLAT SAND
COLOSSUS 2 - C	LIEPAJA, LATVIA	2008	20-25/4	SAND RIPPLES, FLAT SAND, BOULDERS
CATHARSIS 1	PALMARIA, ITALY	2009	11/3-6/4	CLAY, FLAT SAND
CATHARSIS 2	ELBA, ITALY	2009	5-20/10	FLAT SAND, BOULDERS, POSIDONIA
AMICA	TELLARO, ITALY	2010	14/5-11/6	FLAT SAND
ARISE 1	LA SPEZIA, ITALY	2011	2/5-1/6	SAND RIPPLES, FLAT SAND
ARISE 2	ELBA, ITALY	2012	15/10-2/11	FLAT SAND, BOULDERS, POSIDONIA

characteristic highlight-shadow pattern that corresponds to a strong signal return followed in range by a very weak return; the highlight is the result of the echo from the mine itself, while the acoustic shadow that is cast is due to the geometry between the mine (and specifically its height above the seafloor) and the grazing angle of the transmitted signal. In practice, the platform on which the sonar data is usually collected is an autonomous underwater vehicle (AUV).

A. Overview

The performance of the proposed classification algorithm is evaluated using real measured SAS data collected at sea. All of the data used in this study were collected by the CMRE's SAS-equipped AUV called MUSCLE. The center frequency of the SAS is 300 kHz, and the bandwidth is approximately 60 kHz. The system enables the formation of high-resolution sonar imagery with a theoretical across-track resolution of 1.5 cm and a theoretical along-track resolution of 2.5 cm. The present study spans 29 880 SAS images, one example of which is shown in Fig. 5. From the figure, it can be observed how the underwater environment can undergo dramatic seabed changes—e.g., from flat sand to sand ripples—on small length scales.

More specifically, this study examines target classification performance using data collected during six major sea experiments that were conducted by NURC/CMRE between 2008 and 2012; the experiments were Colossus 2, CATHARSIS 1, CATHARSIS 2, AMiCa, ARISE 1, and ARISE 2. The Colossus 2 experiment comprised three distinct subexperiments in different areas, so this division is also respected here. Collection details about the data sets are given in Table I.

In each sea experiment, different groups of man-made targets meant to simulate underwater mines were laid; the targets comprised cylinders, truncated cones, and wedge-shaped objects. Surveys with the MUSCLE AUV were then performed over the target areas (as well as other areas). In total, the area of seabed spanned by the SAS imagery collected during the six sea experiments was approximately 160 km^2. The vast majority of the seabed was flat, benign sand, or clay, but targets were also located in sand ripple fields and posidonia (seagrass). It is this environmental variation for which the proposed classification framework attempts to compensate.

The integral-image-based detection algorithm described in [22] was applied as a prescreener to all of the aforementioned SAS images produced from the surveys. This detector is a cascaded approach with three stages—shadow detection, ripple detection and rejection, and highlight detection—that operate on progressively smaller portions of an image. Binary classification discriminating laid targets from clutter was then to be performed for all alarms generated from this detection stage. Details of the data sets after this detection stage are given in Table II, from which the severe class imbalance of the task can be noted.

TABLE II
DETAILS OF DATA SETS AFTER DETECTION STAGE

SEA EXPERIMENT - AREA	CODE	NUMBER OF TARGETS	NUMBER OF CLUTTER
COLOSSUS 2 - D	COL2D	181	6743
COLOSSUS 2 - B	COL2B	176	12700
COLOSSUS 2 - C	COL2C	128	43364
CATHARSIS 1	CAT1	32	4292
CATHARSIS 2	CAT2	184	26272
AMICA	AMI1	503	20164
ARISE 1	ARI1	142	18684
ARISE 2	ARI2	231	34379

B. Feature Extraction

To facilitate classification, a set of 27 object features was extracted for each alarm. In addition, two environment meta-features were also extracted for each alarm.

1) Environment Meta-Features: The proposed classification framework has several characteristics that are attractive for the underwater mine classification problem. In particular, because the environmental information (i.e., the meta-features) can be extracted directly from the SAS imagery, no special *in situ* measurements (e.g., core samples) must be taken at sea during a survey.

This work employs two environment meta-features that measure the anisotropy and the complexity of the seabed. Essentially, these features are computed via a family of 2-D Haar-like filters, rotated at different angles, to measure directional highlights and shadows in a sonar image. The feature

extraction process that we follow has already been described in detail in [23], so we do not repeat it here.

The anisotropy feature measures the *variation* of the filter responses with an image in different directions. For the SAS images, this effectively quantifies the degree to which the seabed exhibits directional features. For a benign seabed, the anisotropy should be low; for seabed characterized by sand ripples, which appear in SAS images as a series of alternating bands of highlights and shadows (cf., Fig. 5), the anisotropy should be high.

The complexity feature measures the average of filter responses with an image in different directions. For the SAS images, this effectively quantifies the degree to which the seabed contains objects or other sources of irregularity. For a benign seabed, the complexity should be low; for a seabed with irregular variation, the complexity should be high.

To ensure that the presence of the object on the seabed does not skew the environment features, the responses of the alarms are masked prior to the feature calculations. As a result, these environment features should be independent of the object around which they are computed. That is, for a specific patch of seabed, the environment features would ideally always be the same, regardless of the object present.

Next, we provide some evidence of the ability of these features to capture the characteristics of the seabed. Specifically, we show the distribution of these features in three distinct environments: 1) on flat seabed; 2) on seabed characterized by sand ripples; and 3) on the boundary between flat and rippled seabed. In all three cases, the features are extracted from the area of seabed around the same target shape, namely, a truncated cone. For each case, multiple views of the object were obtained from various aspects. One example SAS image from each of the three environments is shown in Fig. 6(a)–(c), with the distribution of the features shown in Fig. 6(d).

The results in Fig. 6(d) suggest that the anisotropy feature is indeed useful for indicating whether seabed is characterized by ripples. The variance of the features extracted for the same object at a given location can be largely attributed to the variation of the images at different views (i.e., aspects and ranges). For example, the rippled seabed can appear nearly flat when the sonar travels parallel to the crests of the ripple field [24]. However, reducing this variance further by refining the feature extraction may be feasible and will be examined in the future.

2) Object Features: For each alarm flagged in the detection stage, a set of 27 object features is also extracted. These features aim to capture various characteristics of the object in order to permit discrimination between targets and clutter in the classification stage. The features are computed using a 5 m $\times$ 10 m SAS image chip around the detection position; example chips were shown in Fig. 6. Ideally, the object features that are extracted to represent each alarm should be invariant to the environment, processing, vehicle motion, and other related factors. However, in practice, it is challenging to create such features. Because object feature extraction is not the primary focus of this work, the list of the 27 object features used in these experiments is withheld until the Appendix.

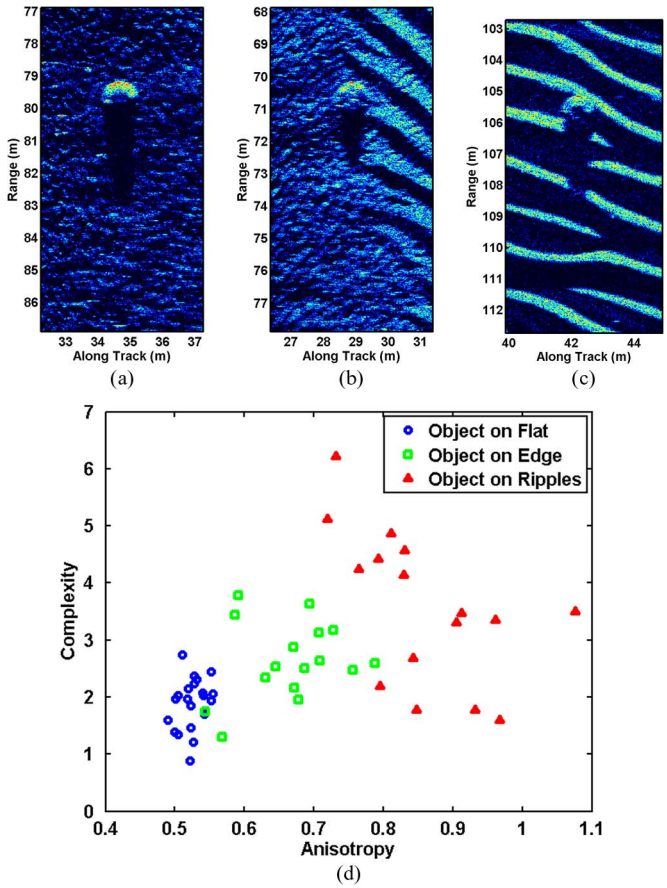

Fig. 6. Example SAS image chip of a truncated-cone-shaped object (a) on flat seabed, (b) on the boundary between flat seabed and ripples, and (c) on seabed characterized by sand ripples; (d) the distribution of the two environment features extracted from the various views of this target shape at these three locations.

IV. Experiments

A. Experimental Setup

Feature extraction was performed for all of the detection alarms from the eight sites. The resulting data were then collected to form eight distinct data sets, one per site. (Ground truth information indicating the true class, target or clutter, of each alarm was also possessed.) A series of eight experiments was then conducted. Specifically, each data set was treated once as unlabeled test data for which prediction was to be performed. For a given test data set, the data from the remaining seven sites were pooled and treated as labeled training data available for learning a classifier.

For each of the eight experiments, the following procedure was carried out. A subset of 100 data points was selected from the training data pool randomly, with only these data used to learn the classifier. Because of the significant class imbalance that exists in this problem (cf., Table II), stratified sampling was used to ensure that both classes were sufficiently represented in the training data (specifically, that at least 20 data points from the target class were present). Predictions were then made on all of the test data.

This process was repeated 100 times, with a new random subset of training data selected for each trial. To ensure fair comparisons, in each trial, every method receives the same

set of training data. The resulting classification performance presented in this section is the mean result from averaging over these 100 trials. This training procedure is similar (though not equivalent) to bootstrap aggregation, or "bagging," which improves robustness [25].

Randomly selecting the training data in repeated trials is beneficial for analysis and algorithm comparison because it allows us to draw more statistically significant conclusions (since the results are based on multiple different training data sets). Doing so also avoids dealing with prohibitively large data sets and their attendant computational issues (e.g., insufficient computer memory). However, future work will explore "big data" techniques to enable training with either all available data or the most informative subset of data [26], [27].

Classification performance is compared for six algorithms, including the three considered in Section II-C: 1) a standard classifier constructed on the 27-D feature data (denoted "Standard RVM"); 2) a classifier constructed on the augmented 29-D feature space that includes the meta-features (denoted "Augmented FS"); and 3) the proposed method. Of the three additional methods, one simply uses an alarm's detection score [22] from the prescreening stage as its final classification prediction. The other two methods use the nonparametric Bayesian context-dependent classification framework introduced by Ratto *et al.* [16]; the *generative* version [15] learns the necessary environment ("context") models using only the environmental meta-features, whereas the *discriminative* version [15], [16] also exploits class labels in the process. The modified RVM employed in Section II-C is used as the (base) classifier for all of the methods.

The detection method provides a baseline measure of performance from which the value of including a classification stage can be quantified. The standard classifier method is included to provide a baseline for the case in which the contextual environmental information is ignored completely. The augmented feature space method shows the change in performance when the environmental information is simply blindly added to the set of object features; this helps demonstrate that performance gains from using the proposed approach are not due to simply having the additional feature available but specifically in *how* that auxiliary information is exploited. The two versions of the framework by Ratto *et al.* provide a comparison to the current state-of-the-art in context-based classification.

The data in the experiments were normalized such that the training data were zero mean and unit standard deviation in each feature dimension. This allows the RVM feature weight magnitudes to be interpreted in terms of relative importance. As is standard practice, a weight associated with a bias term is also included in each RVM classifier, with this enabled by prepending a "1" to each data point's feature vector $\boldsymbol{x}_i \leftarrow [1 \quad \boldsymbol{x}_i^\top]^\top$.

B. Summary Results

Classification performance is presented for each data set in the form of ROC curves and the area under the ROC curve (AUC) [28], which provides a convenient quantitative summary measure of performance. The mean ROC curves, averaged over the 100 random trials, are shown in Figs. 7–14 (a logarithmic scale is used for the probability of false alarm to more easily contrast the performance of the methods). The associated AUC values are shown in Table III, where nonitalicized entries indicate that the improvement of the proposed method's values over the competing methods' values is statistically significant (i.e., $z < -1.96$) according to a Wilcoxon signed-ranks test [29]. As can be seen from the figures and table, the proposed method consistently performed favorably to the competing methods, including the state-of-the-art approaches by Ratto *et al.* Notably, the most significant gains in performance, vis-à-vis the standard classifier that ignores environmental information, were obtained for the challenging Colossus 2C data set, which was comprised of data with substantially different contexts (i.e., seabed characterized by both flat sand and sand ripples).

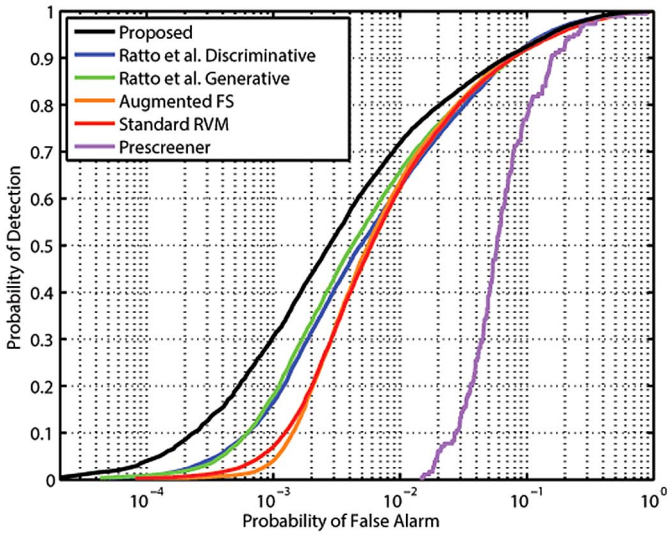

Fig. 7. Classification results when testing on data from the Colossus 2D site after training on data from the other seven sites.

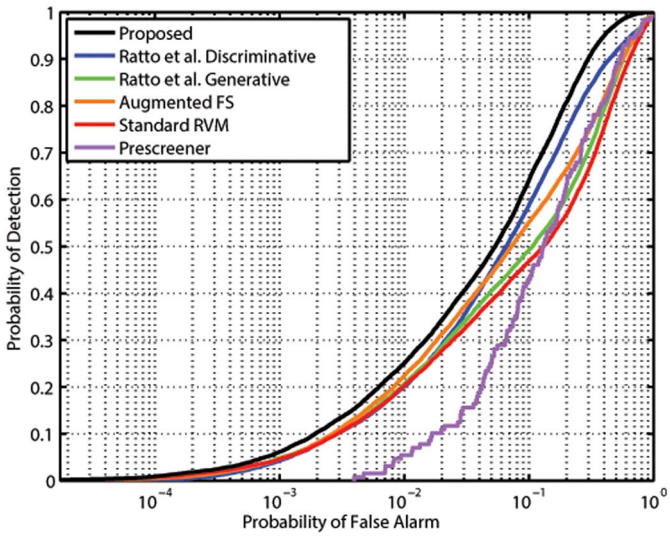

Fig. 8. Classification results when testing on data from the Colossus 2C site after training on data from the other seven sites.

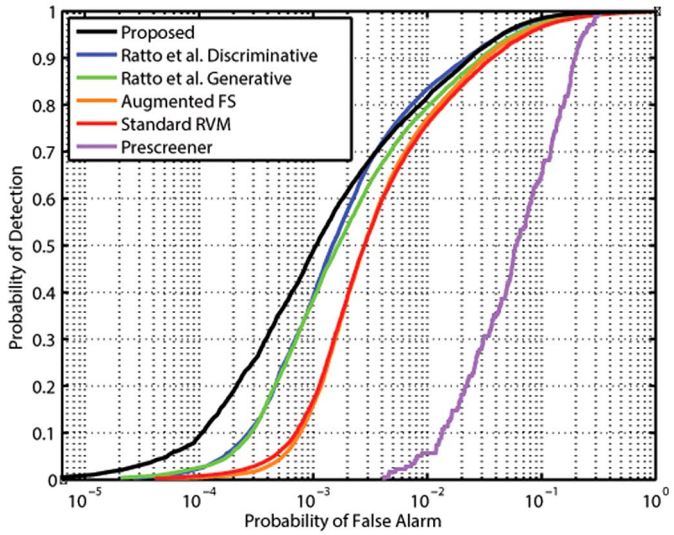

Fig. 9. Classification results when testing on data from the Colossus 2B site after training on data from the other seven sites.

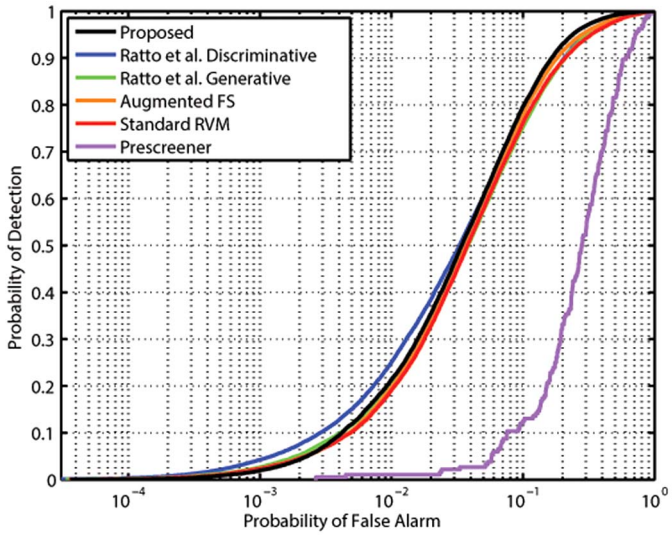

Fig. 11. Classification results when testing on data from the CATHARSIS 2 site after training on data from the other seven sites.

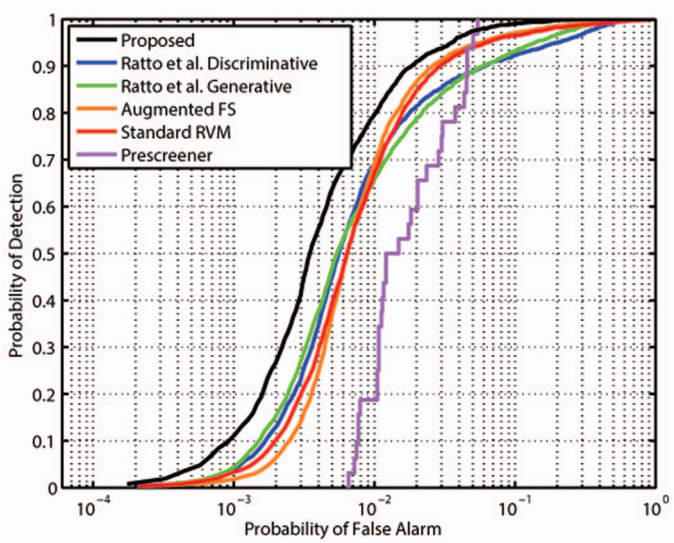

Fig. 10. Classification results when testing on data from the CATHARSIS 1 site after training on data from the other seven sites.

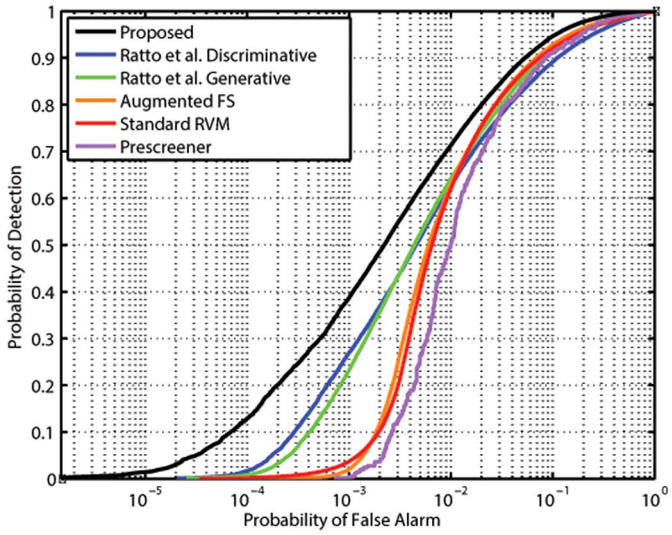

Fig. 12. Classification results when testing on data from the AMiCa site after training on data from the other seven sites.

C. Ranking for At-Sea Contact Reinspection

The underwater mine countermeasures community is particularly interested in conducting completely autonomous operations with sonar-equipped AUVs. The *in situ* processing, analysis, and survey route adaptation using through-the-sensor data has already been demonstrated at sea [30]. The next goal is to enable in-mission survey adaptation to immediately reinspect objects in order to obtain additional views from different aspects [31] to improve classification confidence and further aid in disposal operations. However, resource constraints at sea, such as AUV battery life, mean that only a subset of detected objects can be reinspected. Therefore, it is of vital importance that the list of contacts to be reinspected contains as many actual targets as possible. For this reason, the classifier predictions corresponding to the highest probability of being a mine take on increased significance.

To demonstrate the value of the proposed approach in this regard, the mean number of targets in the set of 100 objects with the highest predicted probability of being a mine for each classification approach is shown in Table IV (the mean is from averaging across the 100 trials). As before, the improvement of the proposed method's values over the competing methods' values is statistically significant (i.e., $z < -1.96$) according to a Wilcoxon signed-ranks test for all nonitalicized entries. As can be observed from the table, the proposed method consistently assigns the highest probabilities to mines, an indication that the method's classification predictions would be a valuable ranking mechanism for *in situ* contact reinspection.

D. Feature Importance

The use of the RVM with a "direct kernel" for the proposed method means that the learned classifier weights indicate the

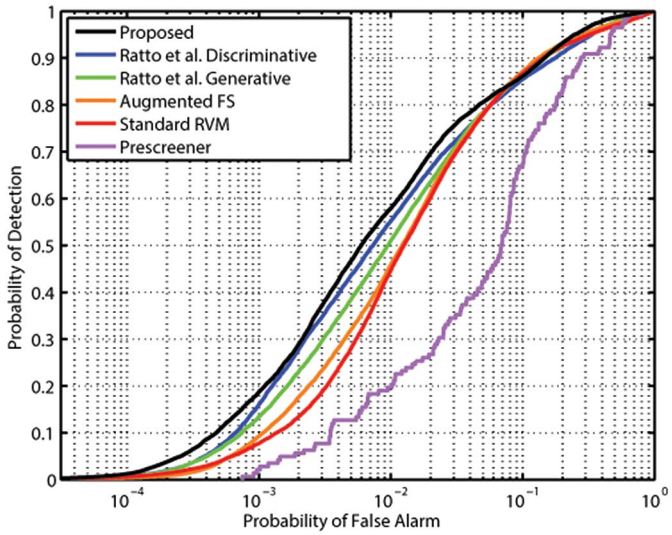

Fig. 13. Classification results when testing on data from the ARISE 1 site after training on data from the other seven sites.

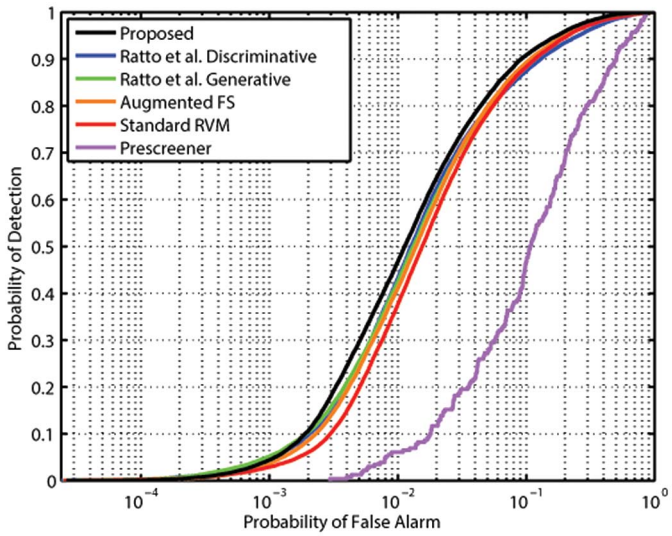

Fig. 14. Classification results when testing on data from the ARISE 2 site after training on data from the other seven sites.

importance of each feature. Over the 100 trials conducted for each of the 8 data sets, thousands of classifiers were learned by the proposed method. Fig. 15 shows the proportion of classifiers (pooling from all experiments conducted) for which each feature was found to be relevant. As can be seen from the figure, features 9, 2, 5, 6, 16, 18, and 1 were relevant in more than a third of the classifiers learned (the interested practitioner can refer to the Appendix for feature descriptions).

However, of more significance is how the *environment* impacted which features were relevant. To understand this, Fig. 16 shows the environmental meta-feature values associated with the classifiers for which each feature was found to be relevant. This figure provides an overall snapshot of how the environment influences the resulting classifiers (the nonuniformity of cases in the upper-left plot is due to the particular meta-feature values of data points in the training sets). For instance, it can be seen that feature 9 is almost always relevant, regardless of the environmental conditions. However, feature 2 is more likely to be relevant only when the anisotropy is low. Features 5 and 6 tend to be more relevant when the complexity is low. These results make sense when one considers what each feature is computing. For example, a low anisotropy indicates an absence of sand ripples, so the template correlation that corresponds to feature 2 would be expected to be useful in those environments; but in high anisotropy cases, the correlation will be skewed by seabed (background) characterized by sand ripples. In low-complexity cases, the seabed surrounding an object is benign, so features 5 and 6 would be robust and hence salient for discrimination; in high-complexity cases, however, shadows in the background would adversely impact the feature calculations.

The proposed algorithm learns C classifiers, each corresponding to different meta-feature values, with the idea that particular *object* features are more useful for discrimination in certain environments. As a result, the classifier parameters, or weights on the features, should differ in different environments. A representative example of this phenomenon is shown in the modified Hinton diagram [32] in Fig. 17.

From this figure, it can be observed how different features are more important for performing classification in certain environments (also, note that many of the parameters are exactly zero, due to the sparsity-promoting properties of the RVM). In particular, it can be seen how the proposed method's individual classifiers associated with different environment meta-features differ from each other and also from a standard RVM that ignores the environment. For example, feature 9 is very important in the standard RVM but relevant only in certain environments, namely, where complexity is not high. Features 5 and 6 were relevant in environments with both low anisotropy and low complexity, while feature 16 was relevant in high-complexity environments; in contrast, none of those three features was relevant in the standard RVM.

E. Discussion

An obvious way to incorporate the auxiliary environment meta-feature into a classification algorithm would be to simply treat it as an additional feature in what would then be an augmented (object) feature space. The experimental results, however, provided evidence that such an approach is inferior to the special manner in which the environmental information is exploited in the proposed framework. In fact, there is a logical explanation to this: The values of the environmental features are, in theory, completely independent of the object with which they are associated. That is, the features are not capturing any information about the object itself, only the seabed surrounding the object. Moreover, since it is assumed that targets are distributed uniformly across environments, there is a principled argument *against* simply treating the environmental features as additional *object* features.

Some of the object features that we extract are known *a priori* to be sensitive to the characteristics of the seabed surrounding an object. For example, highlight and shadow concentration features are computed by explicitly using the background pixels in their calculations, so different seabeds—i.e., different environments—will certainly impact the resulting feature

TABLE III
MEAN AUC

Method	Test Data Set							
	COL2D	COL2C	COL2B	CAT1	CAT2	AMI1	ARI1	ARI2
Proposed	0.9731	0.8911	0.9898	0.9908	0.9324	0.9780	0.9535	0.9622
Ratto *et al.* Discriminative	*0.9702*	0.8449	*0.9900*	0.9681	0.9225	0.9553	0.9416	0.9500
Ratto *et al.* Generative	0.9689	0.7909	0.9866	0.9730	0.9195	0.9661	0.9449	0.9564
Augmented Feature Space	0.9679	0.8102	0.9836	0.9811	0.9244	0.9672	0.9438	0.9553
Standard RVM	0.9661	0.7650	0.9823	0.9788	0.9166	0.9633	0.9393	0.9523
Prescreener	0.9145	0.7855	0.9124	0.9787	0.6785	0.9611	0.8864	0.8134

TABLE IV
MEAN NUMBER OF TARGETS IN THE SET OF 100 OBJECTS WITH HIGHEST PREDICTED PROBABILITY OF BEING A MINE

Method	Test Data Set							
	COL2D	COL2C	COL2B	CAT1	CAT2	AMI1	ARI1	ARI2
Proposed	84.27	13.22	86.92	28.54	13.94	96.38	49.78	26.69
Ratto *et al.* Discriminative	*77.23*	10.28	*83.13*	25.49	*19.33*	89.50	*47.42*	*24.84*
Ratto *et al.* Generative	79.06	10.36	81.99	24.58	*15.23*	86.11	42.69	*25.83*
Augmented Feature Space	73.01	10.59	72.64	26.87	*13.28*	59.44	36.13	22.58
Standard RVM	72.45	10.10	72.91	26.21	*13.20*	57.01	31.44	17.30
Prescreener	0	0	6	19	1	48	18	0

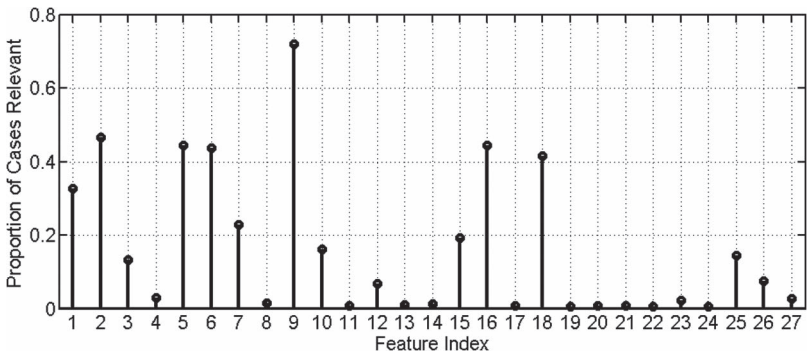

Fig. 15. For the proposed method, the proportion of classifiers learned (pooling from all experiments conducted) for which each feature was found to be relevant.

values. As a result, the features may be salient in benign environments but unreliable when the seabed is characterized by sand ripples. In fact, it is precisely this correlation between object features and the environment that the proposed approach is exploiting. If no dependence exists between the features and the environment, there should be little or no gain from building unique classifiers for different environments. In contrast, the gains achieved by the proposed algorithm in the experiments are likely due to the successful exploitation of this dependence.

The standard RVM method and the approach employing an augmented feature space each build only one classifier. That single classifier is tasked with properly classifying data from (potentially) many different environments. However, there is limited flexibility to do this well in all environments because only one set of classifier weights is available. The result is that the classifier can perform *decently* in many different environments but not *very well* in any one specific environment. (This is similar in spirit to trying to fit a single Gaussian to the data generated by a Gaussian mixture model.) The state-of-the-art methods and the proposed approach enjoy considerably more flexibility because multiple classifiers are learned. As a result, each classifier can be specially tailored to a specific environment, for accurate predictions need to be made essentially on only some—rather than all—data.

The proposed approach performed slightly better than the state-of-the-art methods, and this can perhaps be partially attributed to the fact that the former tended to learn more classifiers than the latter methods (usually 9 for the proposed approach, compared to, on average, 4.2 or 2.4 for the discriminative and generative versions of the state-of-the-art method, respectively). This means that the proposed approach had even more flexibility to tailor its weights to specific environments, whereas the state-of-the-art methods might not have been (effectively) segmenting the environments finely enough. However, it was also observed that the sparsity of the learned classifiers of the proposed algorithm was higher than that of the state-of-the-art approaches (which were all sparser than the single-classifier methods). The higher sparsity, and hence improved

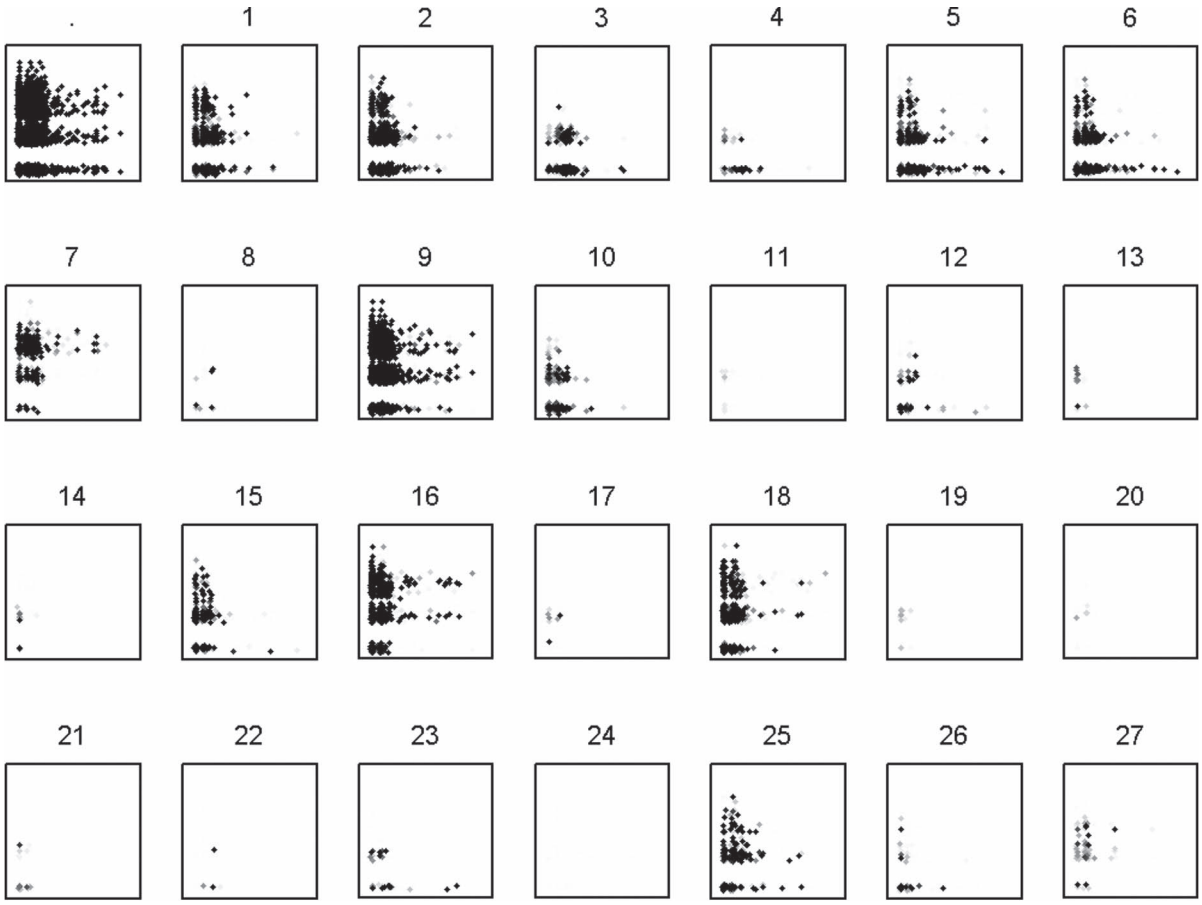

Fig. 16. For the proposed method, feature relevance across all experiments conducted, shown as a function of the environment meta-features. In each subplot, the x-axis corresponds to the anisotropy, and the y-axis corresponds to the complexity. The first (upper left) subplot shows the location of every environment feature pair for which a classifier was constructed in the entire study. Each subsequent subplot shows, for one feature whose index is indicated by the title, the scaled magnitudes of the classifier weight that was learned for each classifier; the magnitude is proportional to the gray-scale value of the marker (black is highest, so cases of irrelevant features are white and hence not visible).

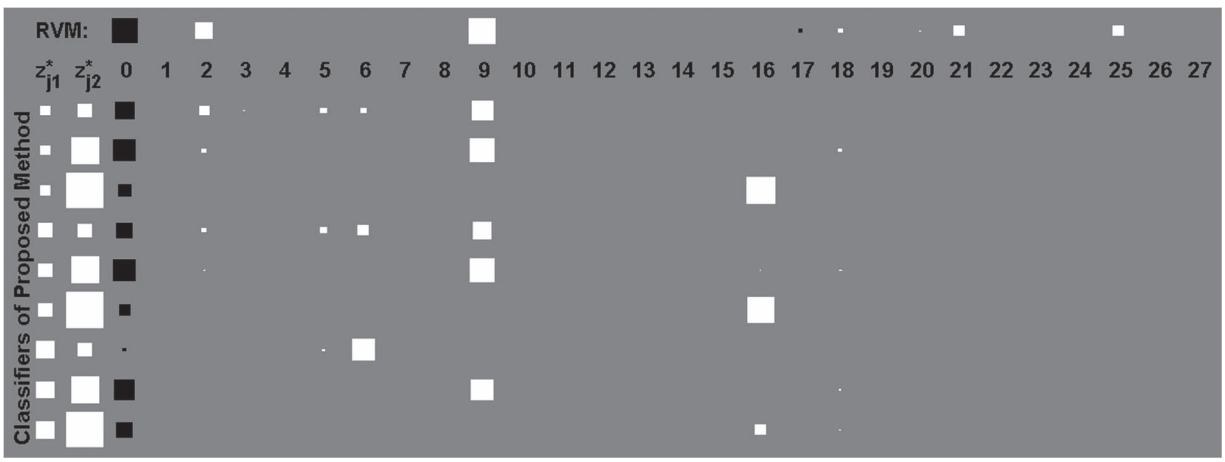

Fig. 17. For one representative experiment, a modified Hinton diagram showing the relative feature weights learned for different classifiers. Each row corresponds to a different classifier; the top row corresponds to a standard RVM, and the nine remaining rows correspond to the nine different classifiers (associated with nine different meta-feature value pairs) constructed under the proposed algorithm. The columns represent the two environment meta-features (seabed anisotropy and complexity, indicated z^*_{j1} and z^*_{j2}, respectively), followed by the RVM bias term (indicated "0") and the 27 object features (indicated "1" through "27"). The size of the square in each column indicates the relative size of the learned classifier weight (or the fixed meta-feature values associated with each classifier); white and black squares indicate positive and negative values, respectively.

generalization ability, may have also contributed to the performance gains. The enhanced sparsity of the proposed approach can likely be attributed to the explicit data-point weighting (based on meta-feature similarity) in the algorithm's classifier learning phase. The sets of most consistently relevant features for the various approaches tended to be similar, with only minor

exceptions (the most notable being that feature 14 was frequently relevant for the discriminative version of Ratto *et al.*).

V. Conclusion

A new classification framework that takes into account environmental information has been proposed. Importantly, the number of classifiers to learn and all other associated model parameters are inferred automatically from the training data by appealing to maximum entropy ideas. The value of the approach was demonstrated experimentally on target classification tasks based on several large SAS data sets collected at sea. In particular, it was shown that—on these data—the proposed approach achieved performance comparable to or better than the state-of-the-art context-based classification methods. The largest gains in performance over the analogous single-classifier approach were observed for data sets with strong environmental diversity, which makes intuitive sense. The value of the proposed method's predictions for guiding AUV-based object reinspection operations at sea was also highlighted. The use of the modified RVM, in which the classifier parameters are weights directly on the features themselves, also helped reveal the features that were most useful for classification in different environments. Insight gleaned from these studies may help lead to the development of better features.

Several directions for future work exist. The experiments conducted here used only a small subset of the total available training data for classifier learning. An efficient way to leverage *all* of the training data is a topic for future work. For instance, the use of landmark data points [33] (to circumvent computer memory limitations) may successfully reduce the size of the training data without incurring significant information loss.

In the context of sonar applications, the development of additional features that capture other salient aspects of the objects is likely necessary to improve performance further. Possible sources of additional information might include object features based on interferometric data [34] and multiview fusion data [35]. The utility of alternative meta-features to characterize the environment should also be examined. To this end, environment meta-features based on modeling the seabed statistics of images [36] or seafloor sediment parameters from backscattering measurements [37]–[39] hold promise. However, it is important that any new environment meta-features devised can still be estimated directly from the data itself.

Appendix

Object Features: For each alarm flagged in the detection stage, a set of 27 object features is extracted from the alarm's SAS chip. These features are described briefly here, where f_i denotes the ith feature.

1) f_1: the score that is the output from the detection algorithm; it is a measure of the highlight (echo) strength of the object.
2) f_2: the maximum correlation of the chip with a basic ternary template that coarsely mimics the highlight-shadow pattern of targets.
3) f_3: similar to f_2 except that the correlation is performed twice with binary templates, one for highlights and one for shadows, separately; the feature value is then the maximum of the two summed correlation maps.
4) f_4: the estimate of the height of the object based on the length of the shadow that it casts, the height of the sonar, and the range of the object from the sonar.
5) f_5: the concentration of the chip's shadow pixels that occur within a prespecified region of the chip expected to contain shadows cast by a target.
6) f_6: the proportion of that same region of the chip expected to contain shadows cast by a target that actually contain shadow pixels.
7) f_7: the concentration of the chip's highlight pixels that occur within a prespecified region of the chip expected to contain highlights due to a target.
8) f_8: the proportion of that same region of the chip expected to contain highlights due to a target that actually contain highlight pixels.
9) f_9: the average of the shadow concentration and highlight concentration features.
10) f_{10}: the maximum raw signal strength of the chip with no image or signal normalization performed.
11) f_{11}: the maximum absolute spatial gradient of the raw signal strength of the chip with no image or signal normalization performed.
12) f_{12}: the mean pixel intensity of a small area around the detection position.
13) f_{13}: the mean pixel intensity of a specified area around the detection position used to extract background statistics.
14) f_{14}: the standard deviation of the pixel intensity values of a specified area around the detection position used to extract background statistics.
15) f_{15}: $f_{15} = (f_{12} - f_{13})/f_{14}$.
16) f_{16}: $f_{16} = (f_{12} - f_{13})/\alpha_1$, where α_1 is the standard deviation of a maximum-response filtering map.
17) f_{17}: the mean intensity of the pixels in a particular orientation associated with highlights.
18) f_{18}: $f_{18} = (f_{12} - f_{17})/\alpha_1$.
19) f_{19}: $f_{19} = f_{16}/f_{18}$.
20) f_{20}: $f_{20} = (f_{13} - f_{17})/f_{14}$.
21) f_{21}: $f_{21} = (\alpha_2 - f_{17})/\alpha_1$, where α_2 is the mean of a maximum-response filtering map.
22) f_{22}: the mean intensity of the pixels in a different particular orientation associated with shadows.
23) f_{23}: $f_{23} = 1 - (f_{22}/f_{12})$.
24) f_{24}: $f_{24} = (f_{13} - f_{22})/f_{14}$.
25) f_{25}: $f_{25} = (\alpha_3 - f_{22})/\alpha_4$, where α_3 and α_4 are the mean and standard deviation, respectively, of a horizontal-response filtering map.
26) f_{26}: $f_{26} = (\alpha_2 - f_{22})/\alpha_1$.
27) f_{27}: $f_{27} = f_{25}/f_{26}$.

Acknowledgment

The authors would like to thank the authors of [16] for enabling performance comparisons by making the code for their algorithms freely available through [40].

REFERENCES

[1] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1999.

[2] V. Myers and J. Fawcett, "A template matching procedure for automatic target recognition in synthetic aperture sonar imagery," *IEEE Signal Process. Lett.*, vol. 17, no. 7, pp. 683–686, Jul. 2010.

[3] H. Midelfart and Ø. Midtgaard, "Robust template matching for object classification," in *Proc. Underwater Acoust. Meas. Conf.*, 2011.

[4] S. Reed, Y. Petillot, and J. Bell, "An automatic approach to the detection and extraction of mine features in sidescan sonar," *IEEE J. Ocean. Eng.*, vol. 28, no. 1, pp. 90–105, Jan. 2003.

[5] R. Fandos and M. Zoubir, "Optimal feature set for automatic detection and classification of underwater objects in SAS images," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 3, pp. 454–468, Jun. 2011.

[6] J. Del Rio Vera, E. Coiras, J. Groen, and B. Evans, "Automatic target recognition in synthetic aperture sonar images based on geometrical feature extraction," *EURASIP J. Adv. Signal Process.*, vol. 2009, p. 14, Jan. 2009.

[7] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in *Proc. Int. Conf. Mach. Learn.*, 2004, p. 114.

[8] M. Sugiyama, M. Krauledat, and K. Muller, "Covariate shift adaptation by importance weighted cross validation," *J. Mach. Learn. Res.*, vol. 8, pp. 985–1005, 2007.

[9] S. Satpal and S. Sarawagi, "Domain adaptation of conditional probability models via feature subsetting," in *Proc. Eur. Conf. Prin. Pract. Knowl. Discov. Databases*, 2007, pp. 224–235.

[10] L. Bruzzone and M. Marconcini, "Domain adaptation problems: A DASVM classification technique and a circular validation strategy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 770–787, May 2010.

[11] S. Yuksel, J. Wilson, and P. Gader, "Twenty years of mixture of experts," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 8, pp. 1177–1193, Aug. 2012.

[12] J. Bolton and P. Gader, "Random set framework for context-based classification with hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 11, pp. 3810–3821, Nov. 2009.

[13] H. Frigui, L. Zhang, and P. Gader, "Context-dependent multisensor fusion and its application to land mine detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 6, pp. 2528–2543, Jun. 2010.

[14] C. Ratto, P. Torrione, and L. Collins, "Exploiting ground-penetrating radar phenomenology in a context-dependent framework for landmine detection and discrimination," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 5, pp. 1689–1700, May 2011.

[15] C. Ratto, "Nonparametric bayesian context learning for buried threat detection," Ph.D. dissertation, Duke University, Durham, NC, USA, 2012.

[16] C. Ratto, K. Morton, L. Collins, and P. Torrione, "Bayesian context-dependent learning for anomaly classification in hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 4, pp. 1969–1981, Apr. 2014.

[17] E. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, no. 4, pp. 620–630, May 1957.

[18] M. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[19] I. Jolliffe, *Principal Component Analysis*. Hoboken, NJ, USA: Wiley, 2005.

[20] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[21] M. Hayes and P. Gough, "Broad-band synthetic aperture sonar," *IEEE J. Ocean. Eng.*, vol. 17, no. 1, pp. 80–94, Jan. 1992.

[22] D. Williams, "On adaptive underwater object detection," in *Proc. Int. Conf. Intell. Robots Syst.*, 2011, pp. 4741–4748.

[23] E. Fakiris, D. Williams, M. Couillard, and W. Fox, "Sea-floor acoustic anisotropy and complexity assessment towards prediction of ATR performance," in *Proc. Int. Conf. Exhib. Underwater Acoust.*, 2013, pp. 1277–1284.

[24] D. Williams, "Bayesian data fusion of multiview synthetic aperture sonar imagery for seabed classification," *IEEE Trans. Image Process.*, vol. 18, no. 6, pp. 1239–1254, Jun. 2009.

[25] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[26] H. Yu, J. Yang, and J. Han, "Classifying large data sets using SVMs with hierarchical clusters," in *Proc. Int. Conf. Knowl. Discov. Data Mining*, 2003, pp. 306–315.

[27] B. Settles, "Active Learning Literature Survey," Univ. Wisconsin, Comput. Sci., Madison, WI, USA, Tech. Rep. 1648, 2009.

[28] J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982.

[29] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bull.*, vol. 1, no. 6, pp. 80–83, Dec. 1945.

[30] D. Williams, A. Vermeij, F. Baralli, J. Groen, and W. Fox, "*In situ* AUV survey adaptation using through-the-sensor sonar data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 2525–2528.

[31] M. Couillard, J. Fawcett, and M. Davison, "Optimizing constrained search patterns for remote mine-hunting vehicles," *IEEE J. Ocean. Eng.*, vol. 37, no. 1, pp. 75–84, Jan. 2012.

[32] S. Nowlan and G. Hinton, "Simplifying neural networks by soft weight-sharing," *Neural Comput.*, vol. 4, no. 4, pp. 473–493, 1992.

[33] J. Silva, J. Marques, and J. Lemos, "Selecting landmark points for sparse manifold learning," *Adv. Neural Inf. Process. Syst.*, vol. 18, pp. 1241–1248, 2006.

[34] T. Sb, S. Synnes, and R. Hansen, "Wideband interferometry in synthetic aperture sonar," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 8, pp. 4450–4459, Aug. 2013.

[35] D. Williams, "SAS and bathymetric data fusion for improved target classification," in *Proc. Int. Conf. Underwater Remote Sens.*, 2012.

[36] J. Cobb, K. Slatton, and G. Dobeck, "A parametric model for characterizing seabed textures in synthetic aperture sonar images," *IEEE J. Ocean. Eng.*, vol. 35, no. 2, pp. 250–266, Apr. 2010.

[37] C. De and B. Chakraborty, "Acoustic characterization of seafloor sediment employing a hybrid method of neural network architecture and fuzzy algorithm," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 743–747, Oct. 2009.

[38] C. De and B. Chakraborty, "Model-based acoustic remote sensing of seafloor characteristics," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 10, pp. 3868–3877, Oct. 2011.

[39] B. Hefner, D. Jackson, A. Ivakin, and G. Wendelboe, "Seafloor characterization using multibeam sonars to support UXO detection," in *Proc. Int. Conf. Exhib. Underwater Acoust.*, 2013, pp. 843–848.

[40] New Folder Consulting PRT: Pattern Recognition and Machine Learning Made Simple 2013. [Online]. Available: http://www.newfolderconsulting.com/prt, New Folder Consulting

David P. Williams received the B.S.E. (*magna cum laude*), M.S., and Ph.D. degrees in electrical and computer engineering from Duke University, Durham, NC, USA, in 2002, 2003, and 2006, respectively.

He was the recipient of a James B. Duke Graduate Fellowship and a National Defense Science and Engineering Graduate Fellowship. Since 2007, he has been with the Centre for Maritime Research and Experimentation (formerly NATO Undersea Research Centre), NATO Science and Technology Organization, La Spezia, Italy. His research interests are in the fields of machine learning, pattern recognition, and mine countermeasures.

Elias Fakiris received the B.S.E. degree in geology and the M.S. and Ph.D. degrees in marine geology and physical oceanography from the University of Patras, Patra, Greece, in 2003, 2005, and 2010, respectively.

He was the recipient of a State Scholarship Foundation of Greece scholarship for his doctorate thesis. Since 2005, he has been a Research Assistant with the University of Patras. In 2012 and 2013, he was a Visiting Scientist with the Centre for Maritime Research and Experimentation, NATO Science and Technology Organization. His research interests are in the fields of geoacoustics, geostatistics, and marine habitat mapping.