

Summary of Progress on SIG Ft. Ord ESTCP DemVal

[Revised Version]

David Williams and Lawrence Carin

Signal Innovations Group
1009 Slater Road, Suite 200
Research Triangle Park, NC 27703

{dpw, lcarin}@siginnovations.com

April 2007

Abstract

We report on progress under an ESTCP demonstration plan dedicated to demonstrating active-learning-based UXO detection on an actual former UXO site (Ft. Ord), using EMI data. In addition to describing the details of the active-learning algorithm, we discuss techniques that were required when applying this method to field data, including a clustering algorithm that plays a key role in properly labeling the data (learning when to label a non-UXO item as UXO-like). The spatially-varying EMI response of each anomaly is first fit using a dipole model, from which each anomaly is characterized in terms of two dipole-moment magnitudes and two resonant frequencies. Information-theoretic active learning is then conducted on all anomalies to determine the labels for the small subset of most-informative anomalies, defined by those that would be most beneficial for classification purposes if the associated label was available. The labels of these most important anomalies are obtained via object excavation. Each anomaly is also characterized by a size feature, obtained by fitting the EMI response to a bivariate Gaussian model. Using this size feature, all anomalies are clustered via a variational Bayesian Gaussian mixture model. Before designing the classifier using the labeled data determined via active learning, the dipole and size features are used to establish which non-UXO items are sufficiently UXO-like that they should be excluded when designing the classifier; we exclude those non-UXO items that are determined to be sufficiently UXO-like, since their inclusion during training would undermine subsequent classifier performance. A kernel matching pursuits (KMP) classifier (using the four dipole-model features) is then constructed. An optional but attractive (based on performance) post-processing step is also considered, this again exploiting the size-feature clustering result. Promising experimental results of our classification algorithm on measured data from the Ft. Ord Seaside UXO site are shown. Several additional experimental results comparing different pairs of algorithms are also shown; specifically, we compare active and non-active labeling procedures, semi-supervised and supervised classifiers, as well as classifiers that account for or ignore concept drift between data sites.

I. INTRODUCTION

In this technical report, we detail the techniques Signal Innovations Group (SIG) has employed for execution of a demonstration of digital geophysics on the Ft. Ord Seaside site. The focus of this demonstration plan was to demonstrate statistical detection of UXO for situations in which one does not have access to *a priori* training data for the current site under test. This problem is addressed via active learning. To explain this technique, we first describe traditional “passive” learning, which constitutes the most widely used form of designing detection and classification algorithms. In passive learning, one assumes access to labeled data with which

to design an algorithm. Labeled data are defined by feature vectors associated with particular anomalies, as well as the associated corresponding label, which specifies whether the feature vector is associated with a UXO or non-UXO item. To constitute this set of labeled data, one must either utilize labeled data characterized (via data collection and excavation) at a previous site, and/or collect a carefully selected set of labeled data by performing measurements of UXO and non-UXO items at the new site of interest. The labeled data is then used to design the classifier. It is important to note that traditional passive learning is characterized as a two-step process: (i) labeled data are acquired from some source, and (ii) this labeled data is used subsequently to design an algorithm. Because these two processes are decoupled, one cannot be assured that the labeled data used are most appropriate for the classification task of interest.

Active learning, by contrast, integrates the collection of labeled data and algorithm design. In active learning, one may assume that no *a priori* labeled data is available. Furthermore, based on sensor data collected at the new site of interest, information-theoretic techniques are used to iteratively define which signatures would be most informative to algorithm design if the associated labels were available. These items are excavated and their associated labels revealed, thereby yielding an adaptive design of the labeled data, directly linked to the site under test [1]. The algorithm automatically determines when a sufficient set of labeled data have been acquired, and therefore, when excavation for the purpose of learning may be terminated. At this point, the labeled data are utilized to build a classifier, and the remaining items are ranked according to the probability of their being UXO, with this constituting the prioritized dig list.

In the active-learning algorithm discussed above, it is assumed that no *a priori* training data is available. In many practical problems, one is likely to have access to prior labeled data from other sites. In such cases, it is desirable to utilize this data, while accounting for the statistical “drift” that may exist between the characteristics of the previous site’s data and the data from the new site of interest. In this scenario, active learning is performed with the objective of learning the relationship between data at the current site and the site characterized by the previously-acquired labeled data. Excavation is then also performed, for the purpose of learning, with this associated algorithm referred to as concept drift [2].

In this report we summarize the demonstration of this technology at the Ft. Ord Seaside site. Important practical considerations must be addressed in such a process. Specifically, the active-learning algorithm seeks labels for signatures that are most informative for learning; consequently,

it seeks labels for those signatures that are most representative of the whole data set. Once it acquires a label for a particular form of signature, it moves on to the next. This presents problems in UXO-sensing tasks for the following reason. If the sensor under test, here EMI, generates a signature that is essentially identical for at least some UXO and non-UXO, then the quality of the resulting algorithm will be strongly dependent on which type of item happened to be excavated. Stated differently, assume we have N different signatures that are essentially identical as seen by the EMI sensor (*i.e.*, they yield essentially the same sensor signature). All of these items will be interpreted as being associated with non-UXO if the active-learning algorithm happens to select a non-UXO; alternatively, all items will be treated as being UXO if the active-learning algorithm happens to select a UXO. Therefore, after implementing the active-learning algorithm at an actual site, a technique must be developed to address this issue. Specifically, that technique must determine which of the labeled data may be trusted in the sense that the data resides in a regime for which the EMI sensor has the ability to successfully classify such data, based on favorable underlying sensor physics. This step is a key and integral component of the active-learning algorithm, and it is addressed here using a clustering algorithm based on variational Bayesian statistics and Gaussian mixture models [3]. As a result of this technique, we determine which non-UXO items are sufficiently UXO-like that they are best not treated as non-UXO items when designing the classifier using the active-learning-determined labeled data.

Data quality is another important issue of concern when implementing active learning or any classification algorithm on field data. In this case, it is important to identify UXO items that become very difficult, if not impossible, to classify because of their proximity to other items (*i.e.*, due to highly overlapping signatures). Of particular concern is the anomaly density, for if items are too proximate the feature-extraction step becomes difficult, and hence classification of individual items is very challenging. One must therefore determine if the data quality associated with a given item is “clean” enough to perform classification, meaning that the item density is sufficiently low in the region of interest. For a site like Ft. Ord, which contains thousands of anomalies, an automated technique is necessary to determine if the data quality of a given anomaly is appropriate for individual-item analysis (*i.e.*, classification). Therefore, as a pre-processing step to the active-learning algorithm, and to all classification algorithms considered here, we have developed a technique that automatically analyzes the data quality, based on the observed data alone. Only those signatures that pass the data-quality-analysis step (implying that

the item of interest is sufficiently separated spatially from surrounding items) are subsequently considered for feature extraction and additional analysis.

A comprehensive summary of our active-learning classification approach is shown in Figure 1. For ease of reference, we include the section numbers of this report in which each component of the algorithm is described.

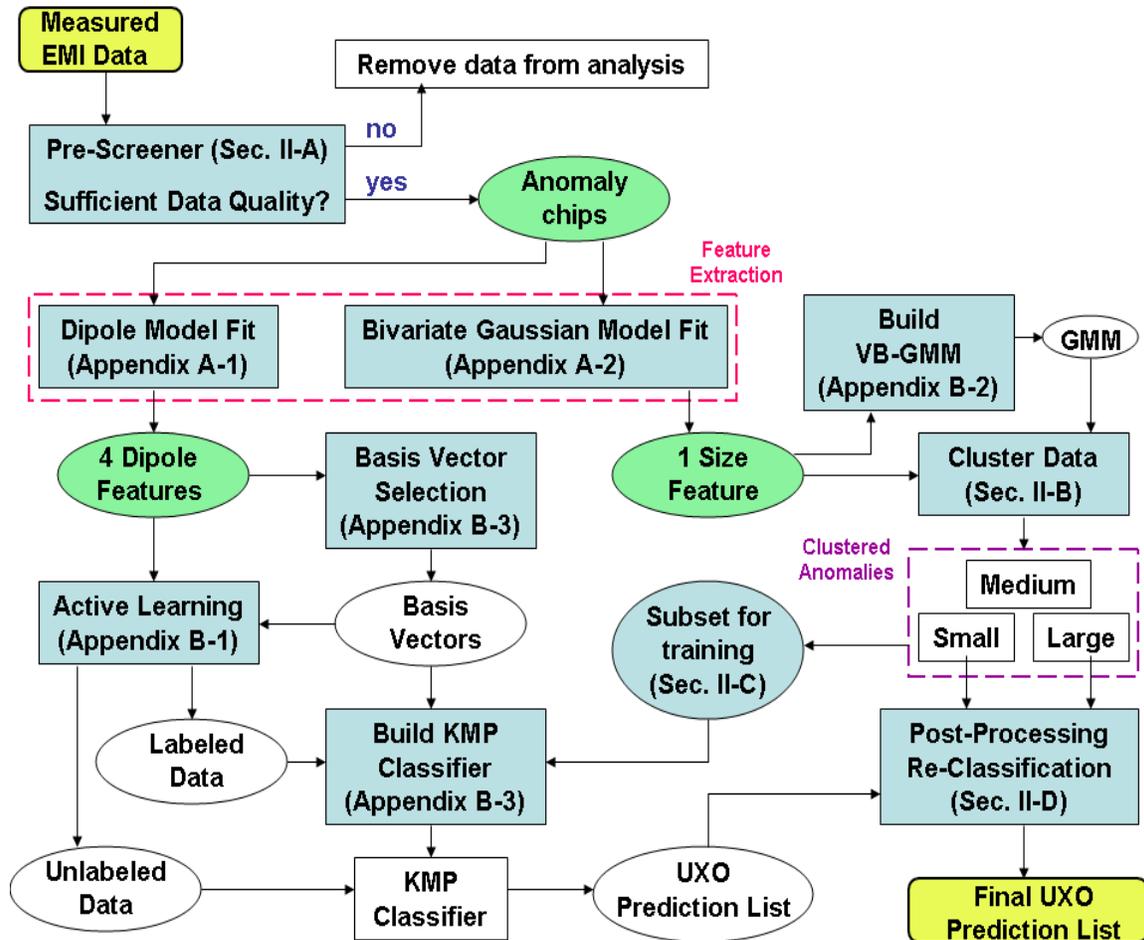


Fig. 1. The main components of the active-learning classification approach presented in this report. Section and appendix numbers in the figure correspond to the locations in this report that describe each algorithm component.

In addition to the main algorithm summarized in Figure 1, we conducted more fundamental experiments that addressed several topics outlined in the demonstration plan. Specifically, in this report we compared the active learning algorithm to a non-active labeling method that randomly chooses the anomalies to label. We also compared a graph-based semi-supervised classification

method to the analogous supervised approach, as outlined in the demonstration plan. Lastly, we examined the importance of accounting for concept drift [2].

The remainder of this report is organized as follows. The overall classification algorithm for UXO detection with no *a priori* labeled data is detailed in Section II. Experimental results of this classification approach as applied to the Ft. Ord Seaside data are shown in Section III. Algorithm performance comparisons — between active and non-active labeling procedures, between semi-supervised and supervised classifiers, and between methods that do and do not account for concept drift — are shown in Section IV. Concluding remarks and directions for future work are given in Section V.

All technical details are provided in the Appendix. In Appendix A we describe the magnetic-dipole and bivariate Gaussian models that are fit to each anomaly’s spatially-varying response; the parameters of these models are used as features in the classification algorithms. Appendix B discusses the technical aspects of the classification algorithm, including active learning, variational Bayesian clustering, and kernel matching pursuits (KMP) classifier design based on the labeled data.

II. OVERALL ACTIVE-LEARNING ALGORITHM FOR ANALYSIS OF FIELD DATA

In this section, we describe in detail the overall active-learning algorithm for analysis of field data. The algorithm is also summarized in Figure 1. The process begins by an initial analysis applied to all measured EMI data, to determine which of the signatures are appropriate for fitting with individual dipole models. Specifically, there are often regions at a given site with sufficient item density that classification of individual items is inappropriate; we have developed a technique to first identify such high-density regions, with the classification algorithms applied subsequently to the remaining data. For signatures that are appropriate for classification, we employ the techniques discussed in Appendices A and B. Specifically, (i) dipole-based feature extraction is applied to the data, (ii) active learning is performed to yield a set of labeled data, (iii) these labeled data are used within KMP to yield a classifier, and (iv) this classifier is then used to quantify the probability that all remaining items are UXO. After all anomalies are so classified based on the result of a KMP classifier, a post-processing step is (optionally) performed. Specifically, certain anomalies are re-classified based on the result of the variational Bayesian GMM clustering discussed in Appendix A.2. This clustering-based classification is based on

the single size feature obtained via the bivariate Gaussian model. The KMP-based classification is based on the four dipole-model parameters (*i.e.*, features). As indicated, the aforementioned post-processing step is not required, but we have found in practice that it significantly reduces the number of items to be excavated, while retaining the same UXO detection probability; in this report we present results determined with and without this post-processing.

A. Preliminaries

The objective in this work is to classify each anomaly at a site as UXO or non-UXO. Here we provide a basic overview of the process we follow to achieve this goal.

First, a simple energy-based pre-screener is applied to the raw sensor data of the entire site, which results in a discrete set of individual anomalies. This initial set of anomalies is then reduced by removing anomalies located in regions for which the anomaly density is not sufficiently low (where anomaly signatures are overlapping). For purposes of dipole model fitting, it is important that each alarm corresponds to a single anomaly. That is, care must be taken to detect alarms characterized by multiple overlapping signatures. To this end, we perform the following prescreening steps. For each of the alarms flagged by the energy detector, a rough estimate of the extent of the alarm is then obtained. This circumscription is performed by locating where the measured data values around an alarm fall below a set threshold. The largest such distance from the alarm center is used as the radius of a circle drawn about the alarm; it is assumed that the signature of the anomaly is wholly contained within this circle. Once the areas of each alarm have been demarcated, we check for overlapping alarm areas. Each set of intersecting alarm areas is merged into a single alarm area and is then removed from further study. Thus, from this preprocessing, the alarms that are retained are only those that do not contain multiple overlapping signatures. This set of anomalies that successfully pass these stages are then compared to the site's anomaly list (supplied *a priori*); the intersection of the set and the list determines the anomalies to consider further. Note that the aforementioned anomaly list corresponds to the items the Corps of Engineers excavated in cleaning up Ft. Ord, and therefore it is only for these items that we may perform the DemVal.

Anomalies characterized by poor data (*e.g.*, sparse spatial sampling of data), as well as anomalies for which signatures of multiple anomalies overlap (discussed above), are considered too difficult vis-à-vis inappropriate for performing classification reliably; these anomalies are not

considered in the subsequent classification step, and in practice would likely be excavated. The remaining anomalies of sufficient data quality are examined further as follows.

Because a predefined list of anomalies is provided, the main purpose of the energy-based prescreener is to eliminate the anomalies on the list for which very low signal energy was present. The final set of anomalies considered was composed of the anomalies that were on both the predefined list and the energy-prescreener list (modulo overlapping anomalies). For example, at the Seaside South site, there were 6,845 anomalies on the predefined anomaly list, while only 1,072 anomalies were contained on both lists and hence analyzed further.

First, each anomaly is fit to the dipole-moment model detailed in Appendix A.1. This process also plays the role of feature extraction, as four of the fit model parameters are retained as features. Next, basis vector selection (using the aforementioned four features) is performed — as outlined in Appendix B.3 — with those anomalies selected as bases also excavated (*i.e.*, labeled). Subsequently, the labels of an additional set of anomalies are labeled via the active learning process, as outlined in Appendix B.1. At the conclusion of the active learning, we possess one set of labeled anomalies and a second set of unlabeled anomalies; using the acquired labeled anomalies, the objective is to design a classifier to determine which of the remaining items are most-probably UXO, with the list of such items constituting the final UXO dig list.

Beyond this point in the process, several different techniques are employed, which we describe in greater detail below.

B. Clustering with Size Feature

In addition to the four dipole-model features that characterize each anomaly, we obtain one size feature by fitting the spatially-varying sensor data to the bivariate Gaussian model in (15) of Appendix A.2. After each anomaly is fit to the bivariate Gaussian model, a variational Bayesian (VB) Gaussian mixture model (GMM) is constituted on the size feature. This GMM will be comprised of K components, with an appropriate value of K determined adaptively based on the data, as discussed in Appendix B.2 ($K = 10$ in this work). The K mixture components in the GMM density function correspond to K clusters of the data by size. One may also choose to apply the VB-GMM algorithm in a recursive fashion on some of the resulting clusters, as we do in this work. This action results in a hierarchical tree structure of clusters (each comprised of fewer data points).

Because the GMM is based on a single feature, one can easily order the mixture components in terms of the magnitude of the (size) features that the mixture contains. Moreover, one can manually assign each component to belong to one of three broad groups: small anomalies, medium-sized anomalies, or large anomalies (based on the properties of the signature). To help guide the manual assignment of mixture components into the three broad groups, one can exploit the labels acquired in the active learning phase (*importantly, note that the active learning is performed on all anomalies that pass the data-quality test, and this clustering analysis is performed subsequently*). Because the active learning is performed *before* the GMM is constructed, each mixture component will most likely contain some labeled data. The mixture components corresponding to the smallest size features *and* which contain (ideally) no labeled UXO can be confidently assigned to the small-anomaly group. Anomalies with size features that are relatively small but that contain at least one labeled UXO would be more properly assigned to the medium-sized anomaly group. In practice, however, these manually-defined divisions between small, medium, and large-sized anomaly groups are not inflexible (*e.g.*, small-anomaly groups with many anomalies may contain some UXO).

This clustering technique is motivated by the observation that the signatures of many anomalies appear to be “blob-like.” With the simple size-feature, coarse classifications (such as dividing anomalies into the small, medium, and large-sized anomaly groups) provide an efficient method to establish anomalies that are clearly *not* UXO-like. When the signatures of many non-UXO appear to be UXO-like, such discriminations are important.

C. Classification with Dipole Features

At most UXO sites, the number of anomalies that are not UXO greatly exceeds the number of anomalies that are UXO. This fact creates a data set with significant class imbalance. A second issue that arises with real UXO data is the problem of overlapping class distributions in feature space. That is, many non-UXO anomalies appear very similar to UXO anomalies in feature space (*i.e.*, of the four dipole-model parameters). As a result, decision boundaries of a classifier cannot be drawn to reliably discriminate between the two classes (UXO and non-UXO). We explicitly address these two major issues in our classifier construction.

The intuitive approach to classifying the unlabeled anomalies would be to construct a classifier using all labeled anomalies (here, as determined via active learning). Such an approach would

ignore the aforementioned insights, however. To address the issue of class overlap in feature space, we begin by using only the anomalies in the medium size feature group as training data (recall the discussion above concerning clustering via the size-based feature). We then exclude those (labeled) *non*-UXO in the medium size feature group whose features closely resemble the features of UXO anomalies. That is, *non*-UXO anomalies from the medium size feature group that look like UXO anomalies in feature space are excluded from the training data set. Specifically, for a given *non*-UXO anomaly, if the absolute value of the difference between any one of its four dipole model features and the average corresponding feature (averaged over the set of labeled UXO) is less than a set threshold, the *non*-UXO anomaly in question is excluded from the training. The exclusion of these anomalies mollifies the issue of class overlap in the training set and yields a desirable *conservative setting* wherein UXO and UXO-like feature vectors are treated as UXO.

To address the issue of class imbalance, we also include in the training set those UXO anomalies from the small-size cluster feature and large-size cluster feature groups whose sizes are most comparable to those in the medium size feature group. That is, the largest (labeled) UXO anomalies from the small-size cluster feature group and the smallest (labeled) UXO anomalies from the large-size cluster feature group are also included in the training data set. The inclusion of these anomalies improves the class balance of UXO and *non*-UXO in the training set.

To summarize, the UXO included in the training data set consist of the UXO in the medium size feature group, and also the UXO most similar in size from the small and large size feature groups. The *non*-UXO included in the training data set consists of the *non*-UXO in the medium size feature group that do not look like UXO in feature space. Using this set of training data, a KMP classifier is constructed. The constructed classifier is then used to obtain each unlabeled anomaly’s probability of being UXO. We emphasize that the pre-processing of the active-learning-determined labeled data, prior to use within the KMP classifier, is an essential tool for utilizing the labeled data properly (*e.g.*, not asking the classifier to distinguish between UXO and *non*-UXO items that are very close in feature space — implying that the underlying sensor physics associated with such items is not appropriate for such distinctions).

It is important to note that our pre-processing of the labeled data does not re-label any of the labels determined via active learning. That is, we do not re-label a UXO-like *non*-UXO item to be treated as UXO. Instead, we remove such *non*-UXO from the training set. Similarly, we

do not re-label small-sized anomalies to be non-UXO, nor do we label large-sized anomalies as UXO. If such re-labeling was performed, the training set would be inundated with new UXO items (because the sensor physics is such that many technically non-UXO items are UXO-like in feature space). Moreover, we also emphasize that the classification is performed using the physics-based dipole features, with the size-based feature from the Gaussian model only used for clustering the data by size.

The constructed classifier discussed above is used to obtain each unlabeled anomaly's probability of being UXO. Setting a threshold for the classifier probability of being UXO effects the binary classification (UXO or non-UXO) of each unlabeled anomaly. By varying this threshold, a receiver operating characteristic (ROC) curve can be generated. In practice, however, a single operating point (and thus a single threshold) must be selected. We select this operating point by setting the threshold using the following principled cost (risk) analysis.

Define C_{01} and C_{10} to be the cost of falsely declaring a UXO to be non-UXO (*i.e.*, false negative), and the cost of falsely declaring a non-UXO to be UXO (*i.e.*, false positive), respectively. Similarly, define C_{11} and C_{00} to be the cost of correctly declaring a UXO to be UXO (*i.e.*, true positive), and the cost of correctly declaring a non-UXO to be non-UXO (*i.e.*, true negative), respectively. Define $p(y = 1|\mathbf{x})$ and $p(y = -1|\mathbf{x})$ to be the probability of a given signature (anomaly) \mathbf{x} being classified as UXO or non-UXO, respectively. The expected cost (risk) of declaring a given signature to be UXO or non-UXO is

$$R_1(\mathbf{x}) = C_{11}p(y = 1|\mathbf{x}) + C_{10}p(y = -1|\mathbf{x}) \quad (1)$$

$$R_0(\mathbf{x}) = C_{00}p(y = -1|\mathbf{x}) + C_{01}p(y = 1|\mathbf{x}), \quad (2)$$

respectively. In our work, the cost of a correct declaration is set to zero (*i.e.*, $C_{11} = 0$ and $C_{00} = 0$), and the cost of a false positive is set to unity (*i.e.*, $C_{10} = 1$). As the cost of a false negative increases, the risk associated with leaving UXO in the ground increases. In practice, a decision-maker decides on the appropriate value for this C_{01} . Varying the value of C_{01} results in the generation of an ROC curve, based on the rules

$$\text{Declare Non-UXO if } \frac{p(y = -1|\mathbf{x})}{p(y = 1|\mathbf{x})} > C_{01} \quad (3)$$

$$\text{Declare UXO if } \frac{p(y = -1|\mathbf{x})}{p(y = 1|\mathbf{x})} < C_{01}. \quad (4)$$

In this work, we choose the operating point of the ROC curve that corresponds to $C_{01} = 100$, meaning the cost of declaring a UXO as non-UXO is one hundred times more costly than declaring a non-UXO as UXO.

D. Optional Post-Processing: Classification with Size Feature

The classifier trained using the active-learning-determined labeled data (and handling the labeled data as discussed above) is implicitly asked to classify anomalies from a wide range of sizes. We note that at most sites there is typically far more non-UXO than UXO, and there are many small non-UXO items. In addition, the sensor physics is unlikely to be sufficient to distinguish the tiny UXO (*e.g.*, grenades) from small clutter. Therefore, as part of an optional post-processing step, one may move all items deemed to be very small (recall the aforementioned size-clustering step) to the end of the UXO dig list. Further, in practice most large anomalies will be excavated in any case. Therefore, all large anomalies, as determined via the size-based clustering, may be moved to the top of the UXO dig list. We present results below with and without this post-processing step, to examine its utility (in the blind test this post processing was used, but we can *a posteriori* examine what our performance would have been if it had not been used).

III. RESULTS

We evaluated the active-learning classification algorithm at the Ft. Ord Seaside site, which was divided into two distinct data sets, Seaside North and Seaside South. The classification analysis was performed using the EMI sensor data that was collected at both sites. This data was collected using a Geonics EM61 MK2, using both the top and bottom coils. The EM61 MK2 was used in the wheel mode, and used differential GPS for positioning. The line spacing was approximately 0.6m, and the point density along track was approximately 15 points per meter. The UXO types at the sites were 37mm projectiles, hand grenades (M1A1, M69, and MKII), 40mm M781, 75mm projectile, trench mortar, trench mortar fuses, 3in. Stokes mortar, 3in. Stokes mortar fuses, 4in. Stokes mortar, and 4in. Stokes mortar fuses.

The classification procedure we employed in these experiments has been outlined in Section II, with technical details provided in the Appendices.

From the pre-screener stage of the algorithm, there were 1,072 and 1,203 anomalies of sufficient quality (as defined previously) present at the Seaside South and Seaside North sites, respectively. After dipole-model fitting (*i.e.*, feature extraction), basis selection, active learning, and the bivariate Gaussian model fitting were performed, the data were segmented into groups via the VB-GMM algorithm. The result of the clustering based on the size-feature is shown in Table I. The results of the active learning, at the Seaside North site, in terms of the size-feature are shown in Table II; a subset of the labeled data is used to build a classifier based on the discussion in Section II-C. The distributions of the features for UXO and for non-UXO at the Seaside North and Seaside South sites are shown in Figures 2 and 4, respectively. These figures demonstrate the lack of separability between the UXO and non-UXO groups in feature space. The distributions of the features for anomalies in the medium-sized group (for which classification was attempted) at the Seaside North site are shown in Figure 3. As can be seen from this figure, discrimination within this subset (of medium-sized anomalies) is more feasible.

TABLE I

NUMBER OF ANOMALIES IN EACH CLUSTER, BASED ON SIZE FEATURE.

	SEASIDE SOUTH	SEASIDE NORTH
TOTAL ANOMALIES	1,072	1,203
SMALL-SIZED ANOMALIES	489	488
MEDIUM-SIZED ANOMALIES	116	386
LARGE-SIZED ANOMALIES	467	329

TABLE II

SEASIDE NORTH ANOMALIES.

	TOTAL NUMBER	NUMBER LABELED VIA ACTIVE LEARNING	NUMBER USED IN CLASSIFIER TRAINING
SMALL-SIZED ANOMALIES	488	208	1
MEDIUM-SIZED ANOMALIES	386	164	52
LARGE-SIZED ANOMALIES	329	108	6

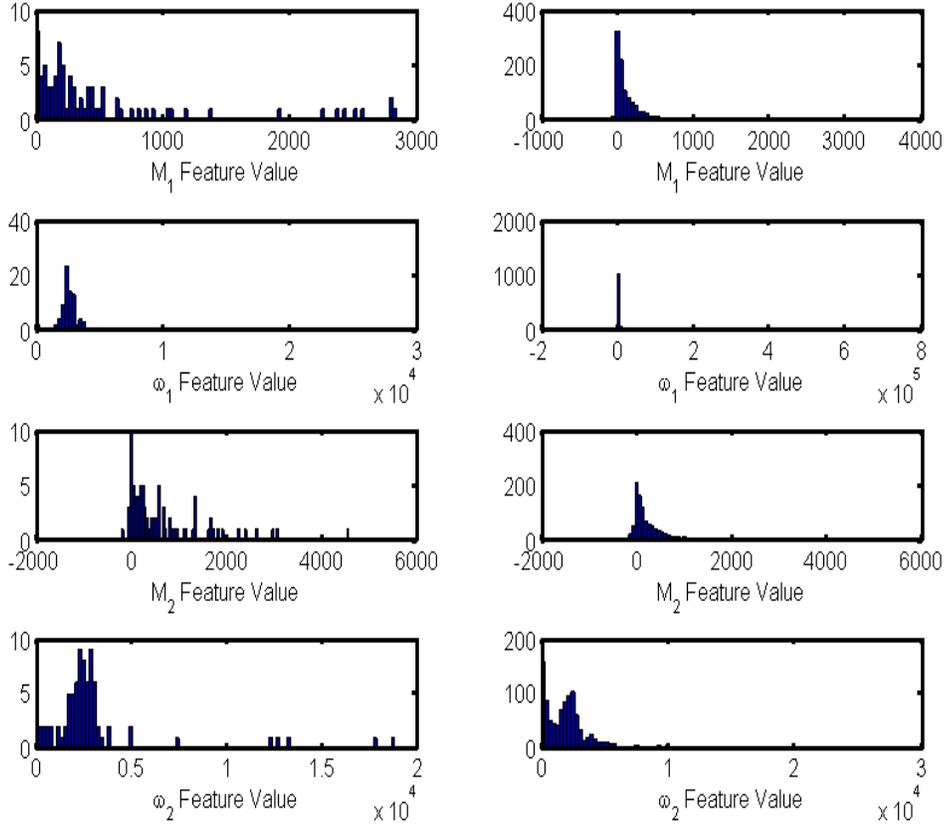


Fig. 2. Distributions of each of the four dipole-model features at the Seaside North site. The vertical axes show the number of occurrences for each feature value (bin). The figures on the left side are for UXO, while the figures on the right side are for non-UXO.

The tree-structured clustering (obtained via repeated application of the variational Bayesian GMM) for the Seaside North site is shown in Figure 5. For the sake of clarity and completeness, we shall here explicitly indicate the size-group to which each of the Seaside North site clusters (shown in Figure 5) are assigned. For this data, the small-sized anomaly group is composed of the anomalies in clusters C-8, C-10, C-4, C-3-8, and C-3-2-8. The large-sized anomaly group is composed of the anomalies in clusters C-3-7, C-3-4, and C-3-2-5. The medium-sized anomaly group is composed of the remaining anomalies, in cluster C-3-2-2.

The classification results of the active-learning classifier are shown in Table III. At the Seaside South site, the two unlabeled UXO anomalies misclassified were both small grenades. In Figure 6, some anomaly chips from the Seaside North site are shown; specifically, this figure shows

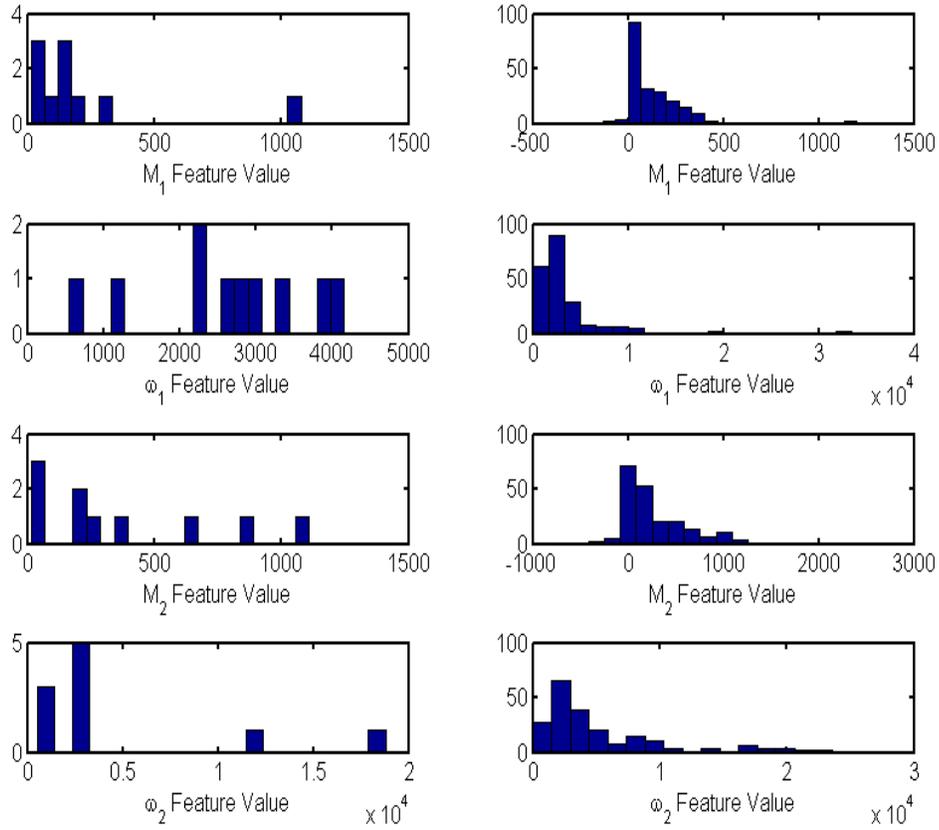


Fig. 3. Distributions of each of the four dipole-model features using only the anomalies in the medium-sized clusters, at the Seaside North site. The vertical axes show the number of occurrences for each feature value (bin). The figures on the left side are for UXO, while the figures on the right side are for non-UXO.

some of the chips of labeled UXO, as well as the five UXO chips that were misclassified by the algorithm. As can be seen from the figure, the five UXO misclassification errors are (apparently) understandable.

Pooling the results of the two Seaside sites, a total of 860 non-UXO anomalies would have been left in the ground, unexcavated. At the same time, by combining the active learning (*i.e.*, labeling excavations) with the classification results, 116 of the total 123 UXO present at the sites would have been successfully detected. This result reflects a 94% probability of detection rate. Thus, this hybrid classification approach successfully achieved a high probability of UXO detection while also leaving many non-UXO buried in the ground. This result would reflect significant savings in terms of remediation costs. This performance achieved with the post-processing step

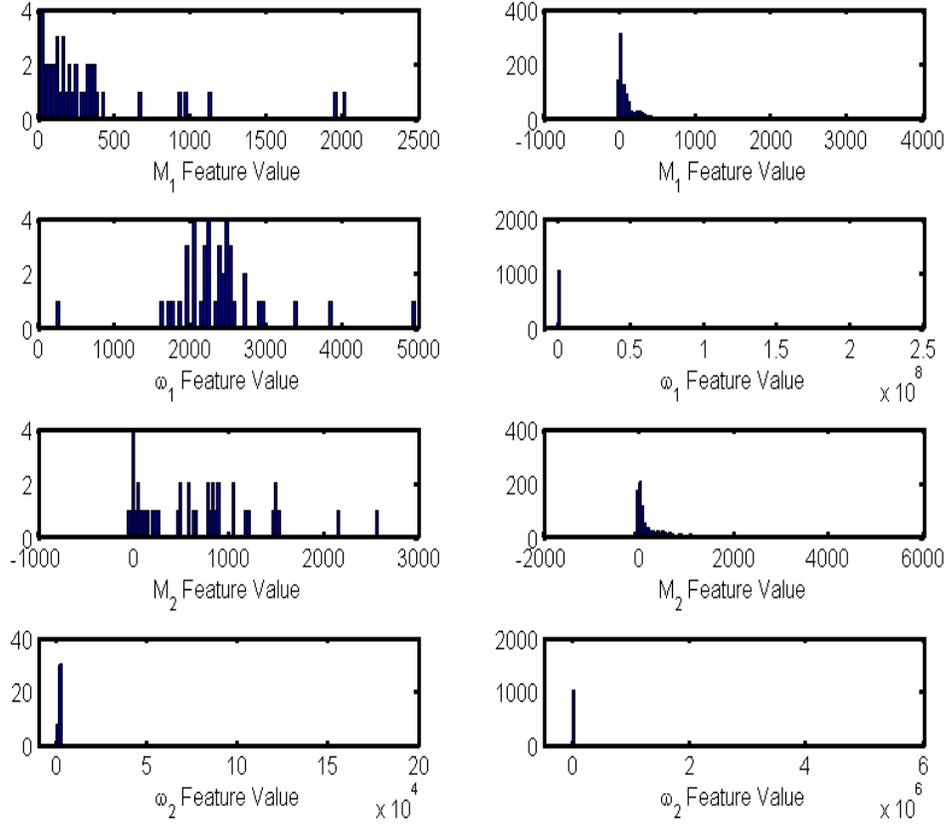


Fig. 4. Distributions of each of the four dipole-model features at the Seaside South site. The vertical axes show the number of occurrences for each feature value (bin). The figures on the left side are for UXO, while the figures on the right side are for non-UXO.

is summarized in Table IV; the performance when the post-processing step is not performed is also shown in the table. Table V shows the performance with the post-processing stage at the Seaside North site when the labeled data (from active learning) is included or excluded in the performance calculations.

To illustrate the value of the components of the active-learning algorithm in an alternative format, we show in Figure 7 performance at the Seaside North site in terms of ROC curves. For the results in the figure, labels were requested for 480 anomalies, which were selected actively. The active-learning algorithm with removal of some of the non-UXO labeled data (green curve and points in the figure) and the (“traditional”) classifiers (cyan and blue curves in the figure) all use the same actively chosen set of labeled data. The green and cyan curves show

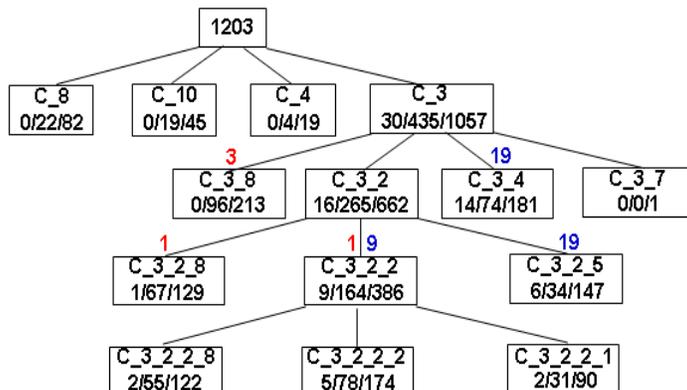


Fig. 5. Tree structure of the clustered Seaside North site data. There are 1,203 total anomalies at the site; 480 of the anomalies were labeled via active learning, 30 of which being UXO. Each box represents a cluster; boxes from which branches (*i.e.*, lines) extend downward are clusters that are segmented (clustered) further again. The first row in a box is the cluster name and number, for description purposes; the second row in a box indicates the number of labeled UXO, the number of labeled objects, and the total number of objects (labeled and unlabeled), that comprise the given cluster. The blue and red numbers above a box indicate the number of unlabeled UXO that were classified correctly or incorrectly by the algorithm, respectively. (For example, in the cluster named “C-3-2-2”, there are 9 labeled UXO out of 164 labeled objects, and 386 total objects. The classification algorithm correctly classified 9 of the unlabeled UXO in this cluster, but misclassified 1 of the unlabeled UXO.) In all, the classification algorithm correctly classified 47 of 52 unlabeled UXO, while correctly classifying (and hence leaving buried) 326 non-UXO.

TABLE III

PERFORMANCE OF HYBRID CLASSIFIER (WITH POST-PROCESSING) AT THE TWO SEASIDE SITES.

	SEASIDE SOUTH	SEASIDE NORTH
TOTAL UNLABELED UXO	41	52
UNLABELED UXO CORRECTLY CLASSIFIED	39	47
UNLABELED NON-UXO CORRECTLY CLASSIFIED	534	326

performance when only a particular subset of the labeled data (described in detail in Section II-C) are used to train the KMP classifier. In contrast, the blue curve shows performance when all of the labeled data are used to train the KMP classifier. The cyan and blue curves correspond to methods that use the KMP classifier to classify all unlabeled anomalies. In contrast, the method corresponding to the green curve also conducts the post-processing stage after classifier predictions for all unlabeled anomalies are made; specifically, the predictions for small-sized and large-sized anomalies — according to the variational Bayesian GMM clustering — are

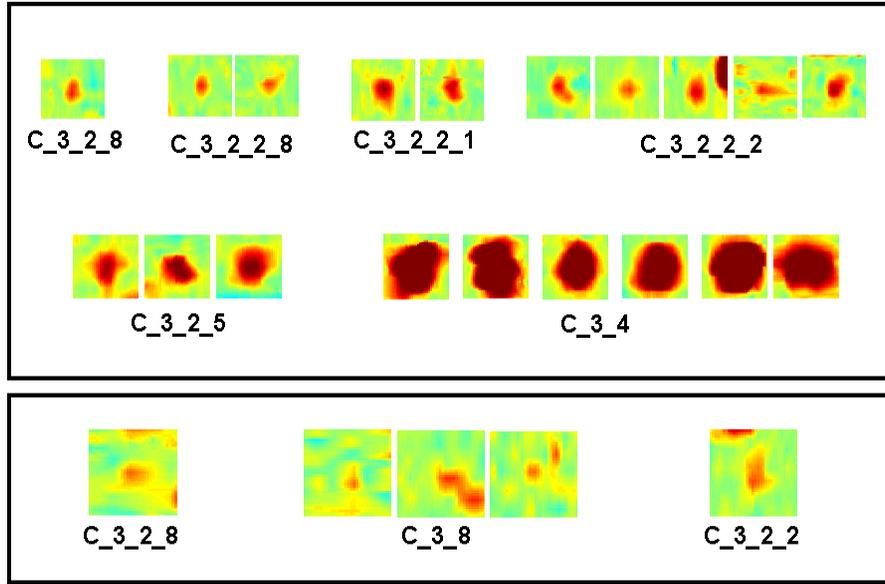


Fig. 6. Chips of UXO separated according to their cluster (from Figure 5). The top box shows some of the labeled UXO, while the bottom box shows the five unlabeled UXO that were incorrectly classified by the algorithm. The algorithm is quite successful in achieving accurate clustering and classification results.

TABLE IV

TOTAL PERFORMANCE OF CLASSIFIER AT THE TWO SEASIDE SITES, WITH AND WITHOUT THE POST-PROCESSING STEP.

	ITEMS WITH USEFUL SIGNATURES	UXO	UXO EXCAVATED	ITEMS NOT EXCAVATED
WITH POST-PROCESSING STEP	2,275	123	116 (94%)	867 (7 UXO MISSED)
WITHOUT POST-PROCESSING STEP	2,275	123	116 (94%)	648 (7 UXO MISSED)

adjusted so that the anomalies in each of these two groups are classified as non-UXO and UXO, respectively. (That is, with the post-processing step, all unlabeled small-size anomalies are declared to be non-UXO, and all unlabeled large-size anomalies are declared to be UXO.) The classifier utilizing a randomly selected set of training data (red curve in the figure) uses the same number (480) of labeled anomalies; the result shown in the figure for this method is the average over 100 independent trials.

As can be seen from Figure 7, the performance of the utilized approach is not a completely continuous curve. This fact is a result of the final post-processing step of classification. All

TABLE V

PERFORMANCE AT SEASIDE NORTH SITE WITH POST-PROCESSING STAGE WHEN THE ACTIVELY-LABELED DATA IS EXCLUDED OR INCLUDED IN THE CALCULATIONS.

	EXCLUDING LABELED DATA	INCLUDING LABELED DATA
PROBABILITY OF DETECTION	0.9038	0.9390
PROBABILITY OF FALSE ALARM	0.5142	0.7092

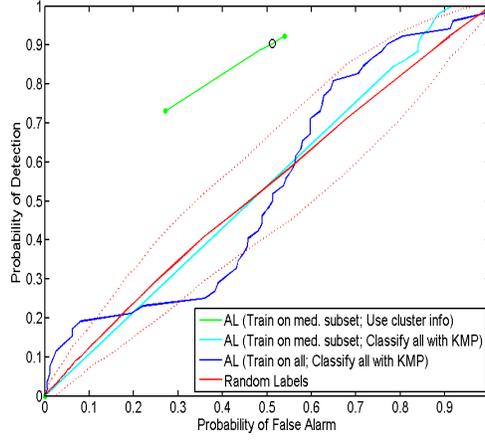


Fig. 7. Algorithm performance comparison on the Seaside North site. The red dashed lines span one standard deviation in each direction about the mean result of the random method, averaged over 100 independent trials. The other three methods use the active learning (AL) algorithm with slight variations. The first two methods (from top to bottom in the insert of this figure) use the same subset of data to train the KMP classifier, but only the first method uses cluster information to adjust class predictions for large-sized and small-sized anomalies. The third method trains the KMP classifier using all labeled data. The first method (green curve) is our method. We choose to operate this hybrid classifier at the black circle; the performance numbers stated in the text correspond to this operating point.

large-sized objects are automatically classified as UXO (responsible for the discontinuity to the right of the point $(0, 0)$), while all small-sized objects are automatically classified as non-UXO (responsible for the discontinuity to the left of the point $(1, 1)$). Between these extremes, the KMP classifier predictions are retained for the remaining medium-sized objects.

Figure 7 displays the significant performance gains that the techniques utilized here realize relative to other classification strategies. By comparing the blue and cyan curves in the figure, it can be seen that training the classifier using only a particular subset of the data does not significantly improve performance. Rather, exploitation of the anomaly size information via the

variational Bayesian GMM clustering is the key to achieving good performance. This observation is supported by comparing the cyan and green curves in the figure. The method corresponding to the green curve employs the post-processing step to adjust its classification predictions on small-sized and large-sized anomalies (based on clustering information), whereas the method corresponding to the cyan curve does not.

We now explain how exploiting the clustering information improves classification performance. When the clustering information is taken into account (post-processing step), the small-sized and large-sized anomalies are automatically classified as non-UXO and UXO, respectively. When this classification adjustment is performed, several anomalies are re-classified. At the Seaside North site, the adjustments to unlabeled data that occur (using the threshold, τ , corresponding to the point at the black circle in Figure 7) are summarized in Table VI. The specific operating point (black circle in Figure 7) was selected as that which reflected the cost of falsely declaring UXO as non-UXO to be 100 times more costly than falsely declaring non-UXO as UXO.

TABLE VI

EFFECTS OF POST-PROCESSING STEP AT SEASIDE NORTH SITE. SMALL AND LARGE ANOMALIES (BASED ON SIZE FEATURE) ARE RE-CLASSIFIED AS NON-UXO AND UXO, RESPECTIVELY. POSITIVE CHANGES ARE CLASSIFICATIONS THAT BECOME CORRECT WITH THE RE-CLASSIFICATION, WHILE NEGATIVE CHANGES ARE CLASSIFICATIONS THAT BECOME INCORRECT.

KMP CLASSIFIER PREDICTION	NON-UXO		UXO	
SIZE FEATURE GROUP	LARGE		SMALL	
TRUE ANOMALY IDENTITY	NON-UXO	UXO	NON-UXO	UXO
CLASSIFICATION CHANGES DUE TO SIZE FEATURE	-18	+4	+237	-4

As can be seen from Table VI, there is no net change in classification of UXO as a result of the clustering-based prediction re-assignment. However, the re-assignment step does reduce the number of false alarms by 219 (*i.e.*, 237-18), meaning 219 additional non-UXO anomalies can be left in the ground unexcavated by exploiting this post-processing step. It is for this reason that the green ROC curve is far superior to the cyan ROC curve in Figure 7.

By addressing the problems of class overlap and class imbalance, and by also exploiting anomaly-size information, the techniques utilized here circumvent limitations from which traditional classifiers would suffer.

IV. ALGORITHM COMPARISONS

In addition to the active-learning classification approach discussed heretofore, we conducted more fundamental experiments that addressed several topics outlined in the DemVal plan. Specifically, we compared our active-learning algorithm, which adaptively selects the anomalies to label, to a non-adaptive labeling method that randomly chooses the anomalies to label. We also compared a graph-based semi-supervised classification method to the analogous supervised approach. Lastly, we examined the impact of concept drift on classification performance. In all of these experiments, only the dipole-model features were employed. We believe that these tests, in addition to the results presented in Section III, completely address all aspects of the DemVal plan.

It should be noted that we sought to isolate the effects of a single variable in each of the experiments that follow. Therefore, rather than including the various components of our main classifier from Section II — such as the use of a particular subset of the labeled data for classifier design, or the post-processing step — the following experiments employ standard learning approaches. This choice permitted us to more effectively discern performance gains directly attributable to the algorithm under study (*i.e.*, active learning, semi-supervised learning, or concept drift). That is, we were able to more accurately evaluate the utility of the various techniques.

A. Active Learning

1) *Theory*: The active-learning algorithm is presented in Appendix B.1, so we do not repeat the discussion here.

2) *Results*: We conducted experiments on both sites, Seaside North and Seaside South, to evaluate the utility of actively selecting the data to label. To this end, we compared classification performance when the labeled data was selected actively and when it was selected randomly. The criterion upon which the active method selects data to label is given in Appendix B.1. In the non-active case, we randomly selected the data points to use as labeled (training) data. A KMP classifier was used for both approaches.

For the Seaside North and Seaside South sites, 77 and 35 anomalies were chosen to be labeled, respectively. These numbers of anomalies were chosen based on a set information gain threshold. (Many fewer anomalies are chosen to be labeled in this set of experiments than in

those conducted in Section III because a lower information-gain threshold was used in the latter study.) We chose the same number of anomalies to label using the non-active (random) methods as were chosen using the active methods. One hundred different trials were conducted for the non-active cases, where each trial used a unique set of labeled data. *For each trial of the non-active cases, it was ensured that at least one of the labeled anomalies was UXO.* In practice, however, it is possible that none of the randomly selected anomalies would be UXO; in such a scenario, classification would be impossible. Therefore, this experimental set-up may unfairly favor the non-active method.

The comparison between the active case and the non-active case (with random labeling selections) is shown in Figure 8. As can be seen from the figure, the active method achieves far superior performance compared to the non-active method. The significant difference between the performances of the two methods can be attributed to the fact that there are relatively few UXO anomalies in the data set. As a result, randomly selecting the anomalies to label will tend to result in very few UXO anomalies chosen to be labeled. In turn, the labeled training set will not be a representative set of the entire data set. It should be noted that this adaptive study we conducted to satisfy the demonstration plan is different from the hybrid classification method discussed in Section II that employed the clustering algorithm.

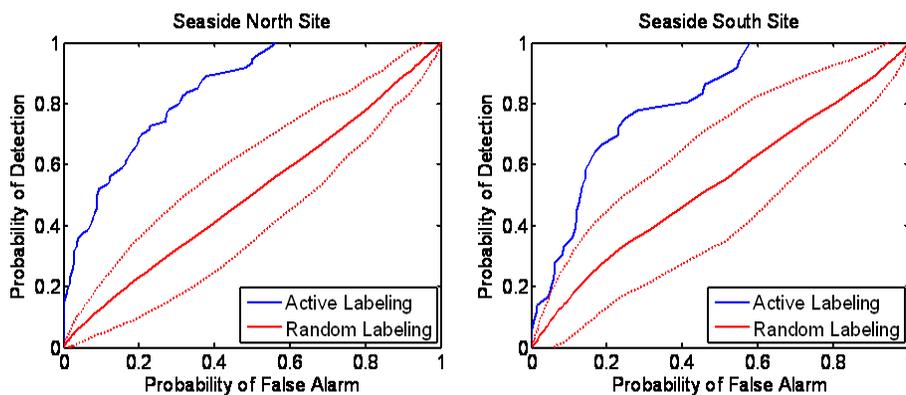


Fig. 8. Comparison between the active learning method for choosing which labels to acquire, and a non-active method that randomly chooses which labels to acquire. The results are for (left) the Seaside North site, and (right) the Seaside South site. The red dashed lines span one standard deviation in each direction about the mean result of the random method, averaged over 100 independent trials.

A comment should also be made regarding the significant difference in performance of the traditional classifiers that employ active learning in Section III and in this section. In the

experiments in Section III, a large number of anomalies were chosen to be labeled, whereas the experiments in this section selected a rather small number of anomalies to be labeled. In these latter experiments, it is less likely that a (somewhat uncommon) non-UXO anomaly that looks like a UXO will be selected to be labeled. It is also less likely that a (rare) UXO anomaly that looks like a non-UXO will be selected to be labeled. As a result, the (actively chosen) labeled training data that is used to construct a classifier in this case should not suffer from severe class-distribution overlap. That is, the labeled training data should be relatively well-separated. In turn, an effective classifier would be built. In the (Section III) experiments in which a substantial portion of the data set was selected to be labeled actively, however, the training data will likely be contradictory in the sense that the class-distributions heavily overlap. Any classifier constructed using such training data will tend to be poor. It is for this reason that the various aspects of the hybrid classification properly treating the labeled data were deemed so vital, and made such a profound impact on performance in Section III.

B. Semi-Supervised Classification

1) *Theory*: Semi-supervised algorithms utilize both labeled and unlabeled data to build a classifier [4], whereas supervised algorithms utilize only labeled data.

In addition to a set of labeled data, $\mathcal{D}_L = \{\mathbf{x}_i, y_i\}_{i=1}^{N_L}$, assume we have a set of unlabeled data $\mathcal{D}_U = \{\mathbf{x}_i\}_{i=N_L+1}^N$ for which the labels are unknown. A kernel function measures the similarity between two data points. Computing the kernel function for every pair of N data points (both labeled and unlabeled) results in the symmetric, positive semidefinite kernel matrix \mathbf{K} . The ij -th element of the kernel matrix — K_{ij} — is a measure of similarity between data points \mathbf{x}_i and \mathbf{x}_j . With \mathbf{D} the diagonal matrix whose ii -th element is given by $D_{ii} = \sum_{j=1}^N K_{ij}$, the graph Laplacian is defined to be

$$\mathbf{\Delta} = \mathbf{D} - \mathbf{K}. \quad (5)$$

A fully connected, undirected graph with vertices $V = \{1, 2, \dots, N\}$ can be summarized by the above kernel matrix \mathbf{K} in the following manner [5]. By assigning one vertex of the graph to each data point, the edge of the graph joining vertices i and j can be represented by the weight K_{ij} . A natural way to measure how much a function $\mathbf{f} = [f_1, \dots, f_N]^T$ defined on V

varies across the graph is by the quantity

$$\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N K_{ij} (f_i - f_j)^2 = \mathbf{f}^T \Delta \mathbf{f}. \quad (6)$$

By defining a Gaussian random field (GRF) on the vertices V ,

$$p(\mathbf{f}) \propto \exp\{(-\lambda/2) \mathbf{f}^T \Delta \mathbf{f}\}, \quad (7)$$

smooth functions \mathbf{f} are deemed more probable. In (7), λ is a positive regularization parameter. If we define $f_i = \mathbf{w}^T \mathbf{x}_i$, then $\mathbf{f} = [f_1, \dots, f_N]^T = \mathbf{X}^T \mathbf{w}$, where the ai -th element of \mathbf{X} corresponds to the a -th feature of the i -th data point. With this choice, $p(\mathbf{f})$ induces a Gaussian prior on \mathbf{w} ,

$$p(\mathbf{f}) = p(\mathbf{w} | \{\mathbf{x}_i\}_{i=1}^N) \propto \exp\{(-\lambda/2) \mathbf{w}^T \mathbf{X} \Delta \mathbf{X}^T \mathbf{w}\} = \exp\{(-\lambda/2) \mathbf{w}^T \mathbf{G} \mathbf{w}\}, \quad (8)$$

with the precision matrix $\mathbf{G} = \mathbf{X} \Delta \mathbf{X}^T$. This formulation encourages “similar” data points to have similar class labels.

The resulting classifier weights \mathbf{w} constructed using this formulation will reflect the influence of the prior.

2) *Results:* We also compared a supervised learning algorithm with a semi-supervised method, as outlined in the demonstration plan, for both the Seaside North and Seaside South sites. We randomly selected 100 data points to use as labeled (training) data at each site, in each of 50 independent trials. A KMP classifier was used in all experiments. The difference between the two approaches is that the semi-supervised method exploits the unlabeled data in the classifier design, whereas the supervised method does not. Specifically, the semi-supervised approach also employs a graph-based prior to influence the construction of the classifier.

The comparison between the supervised and semi-supervised cases is shown in Figure 9. As can be seen from the figure, the semi-supervised approach does not improve performance above that of the supervised method. This result can be attributed to the fact that the distributions of the two classes (UXO and non-UXO) overlap heavily. As a result, the semi-supervised approach cannot successfully exploit the structure of the manifold on which the data lies via the prior.

It should be noted that this study is different from the classification method discussed in Section II, that employed the clustering algorithm. The nature of the latter approach, involving clustering, does not lend itself to a direct comparison with a semi-supervised method.

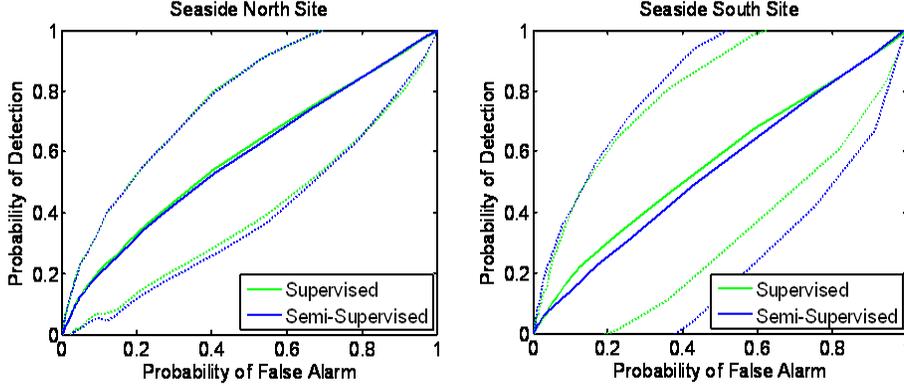


Fig. 9. Comparison between the supervised and semi-supervised learning methods. The results are for (left) the Seaside North site, and (right) the Seaside South site. The dashed lines span one standard deviation in each direction about the mean result of each method, averaged over 50 independent trials.

C. Concept Drift

1) *Theory*: A common assumption of learning algorithms is that the underlying statistics that generate the training and testing data are the same. However, in practice, when training and testing data come from different geographical sites, this assumption may be violated. For example, changes in environmental effects or sensor operating conditions used in the data collection at different sites can create sets of data with different underlying statistics. Concept drift [2] is an algorithm that explicitly accounts for these underlying differences.

In the concept drift algorithm, it is assumed that there are two distinct sets of data. A set of (N^a) *auxiliary* data is assumed to be labeled, available to be used as training data. The second set of data, on which we wish to perform classification, is from a different *primary* data site. Initially, all of the primary data is unlabeled. In order to learn the relationship — and in turn the type of drift — between the the primary and auxiliary data, one (actively) selects N_L^p data points from the primary data set to be labeled.

To account for the differences between the two data sets, one unique auxiliary variable, μ_i , for each auxiliary data point is included in classifier learning. Specifically, the concept drift algorithm [2] employs a logistic regression framework in training on the subset of labeled primary data, \mathcal{D}_L^p , and the set of auxiliary data, \mathcal{D}^a . The log-likelihood function is given by

$$\ell(\mathbf{w}, \boldsymbol{\mu}; \mathcal{D}_L^p, \mathcal{D}^a) = \sum_{i=1}^{N_L^p} \log \sigma(y_i^p \mathbf{w}^T \mathbf{x}_i^p) + \sum_{i=1}^{N^a} \log \sigma(y_i^a \mathbf{w}^T \mathbf{x}_i^a + y_i^a \mu_i), \quad (9)$$

where $y_i \in \{-1, 1\}$ is a label, \mathbf{x}_i is a feature vector (*i.e.*, data point), \mathbf{w} is a vector containing the classifier weights, σ is a sigmoid function, and superscripts indicate primary or auxiliary data sets.

The classifier that accounts for concept drift is learned by simultaneously learning the classifier weights \mathbf{w} and the auxiliary variables $\boldsymbol{\mu}$ via maximization of (9). The auxiliary variables are the key to the algorithm, as they introduce flexibility into the model. Moreover, they implicitly describe the relationship between the data at the primary site and the data at the auxiliary site. Specifically, the auxiliary variables control the participation of each auxiliary data point in learning the classifier \mathbf{w} . When an auxiliary data point is mismatched to the primary data, the value of the corresponding auxiliary variable will be larger, which in turn limits the impact of that data point in learning \mathbf{w} . In this manner, the auxiliary data can be beneficial in learning a classifier for the primary data, despite the underlying differences between the two sites.

2) *Results*: We also conducted two sets of experiments to evaluate the relevance and utility of the concept drift algorithm. In one set of experiments, the Seaside North data set was used as the primary data site at which we wished to perform classification; in the second set of experiments, the primary data site was the Seaside South data set. The Seaside North data set was comprised of 1,203 anomalies, 82 of which were UXO. The Seaside South data set was comprised of 1,072 anomalies, 41 of which were UXO. For all experiments, the Badlands Bombing Range (on the Pine Ridge Reservation in South Dakota) was used as the auxiliary data site, acting as training data for which it was assumed all labels were available. This data set was comprised of 492 anomalies, 57 of which were UXO.

The Badlands data was collected by a pulsed induction array that uses highly-modified Geonics EM-61 sensors. The modifications are intended to make the sensors compatible with vehicular towing and increase the sensitivity to small objects. The Badlands data covers over 150 acres at two range targets, designated BBR1 and BBR2. BBR1 is a highly visible circular target composed of a 500-ft-diameter circular earth berm (3 to 5 ft high), with a cross hair berm inside the circle. This bull's-eye was used primarily as a bombing target by World War II vintage aircraft. The ordnance dropped on this target was primarily M 38 (100 lb.) sand-filled bombs. These sheet metal structures were typically highly deformed by impact, some look like pancakes, some like basketballs and some have lost structural integrity. Almost all ordnance that penetrated lost the box tail fin and burster train components at the surface. The munitions types at Badlands

included M 38 (100 lb.) sand-filled practice bombs, M 57 (250 lb.) practice bombs, 2.25 in. and 2.75 in. rocket bodies and rocket warheads, and ordnance scrap (such as tail fins and casing parts).

Because the models used to extract features from the EM-61 sensor data at the Seaside and Badlands sites were slightly different, the features used in the experiments were limited to those in common. In this case, only two features were common to the data at both sites: the two dipole moments (from dipole model fitting). Therefore, only these two features were used to characterize each anomaly, as the concept drift algorithm requires that the same (nominal) features characterize the data from the primary and auxiliary data sites.

We compare four different methods in this study, each of which employs a logistic regression framework for classification. In one method, a classifier was trained using only the (Badlands Bombing Range) auxiliary data. The second method trained a classifier using only 50 labeled anomalies from the primary data (Seaside North or Seaside South). These 50 anomalies were selected actively, and were also used as the labeled primary data in the remaining two methods. In the third method, a classifier was trained by pooling all of the auxiliary data and the 50 labeled primary data points, treating them equally (*i.e.*, ignoring the drift). The fourth and final method used the same training data as the third method, except the concept drift algorithm was utilized.

The classification results from this study are shown in Figure 10. The ROC curves show the classification performance on the data from the primary sites. As can be seen from Figure 10, the proposed concept drift algorithm improves performance over the competing methods on the Seaside North data, but worsens performance on the Seaside South data. It should also be noted that many fewer anomalies from the Seaside sites were labeled for these experiments, compared to the experiments in Section III. This fact is because the purpose of these concept-drift experiments is to assess the utility of concept drift relative to other approaches *that use the same data*.

V. CONCLUSION

In this classification study, we analyzed the Ft. Ord Seaside UXO site. We conducted several experiments comparing different pairs of algorithms; specifically, we compared active and non-active labeling procedures, semi-supervised and supervised classifiers, as well as classifiers

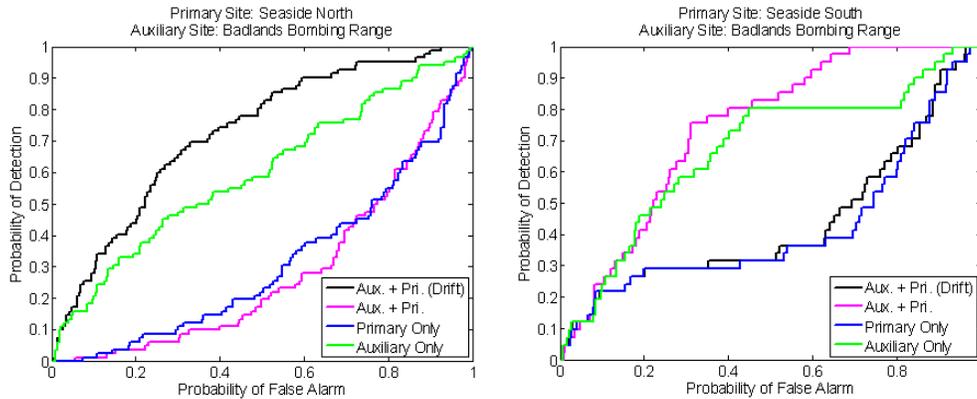


Fig. 10. Comparison of concept drift algorithm to alternative methods. The primary data site is (left) the Seaside North site, and (right) the Seaside South site; in both cases, the auxiliary data site is the Badlands Bombing Range. In each set of experiments, the labels of 50 anomalies from the primary site were actively acquired.

that account for or ignore concept drift between data sites. The performance of the active labeling procedure far exceeded that of the non-active method. The semi-supervised classification approach did not improve performance beyond that of the supervised method, because of the heavy class overlap of the data. The effects of the concept drift algorithm on performance were mixed. At the Seaside North site, the method improved performance over the alternative methods that did not account for the differences between the (primary) Seaside North site and the (auxiliary) Badlands site. However, at the Seaside South site, performance worsened when the concept drift method was used. It should be emphasized that this was an incomplete analysis of concept drift, since the sensors deployed at Ft. Ord and the Badlands were different, and therefore we had to prune the features to the subset that overlapped between the two. We are confident that concept drift will add value when properly used between sites, *i.e.*, when the same sensor is deployed at the multiple sites.

We also utilized an approach to the classification of UXO that employed both a clustering technique and a probabilistic classifier. To perform the clustering aspect of the algorithm, a variational Bayesian Gaussian mixture model (GMM) was employed. This clustering was based on a size-based feature obtained from fitting the responses of anomalies to a bivariate Gaussian model. Active learning and the probabilistic classifier — kernel matching pursuits (KMP) — were both based on four dipole-model features extracted via a (dipole) model-inversion procedure with the measured data. The bivariate Gaussian model and the *time*-domain dipole model were

both new contributions. At the Seaside sites, the employed classification approach successfully detected 116 of 123 UXO anomalies, while correctly classifying (and hence leaving unexcavated) a total of 981 non-UXO anomalies. We also explicitly demonstrated the importance of the algorithm’s final post-processing step of re-classification based on the size feature.

In the future (under proposed SERDP support — not part of this ESTCP project), we hope to experiment with alternative approaches to the classification algorithm considered here. For example, in the approach employed in this work, active learning was performed prior to segmenting the data into clusters. One alternative method would reverse the order of these steps, first clustering the data, and then performing basis vector selection, active learning, and classifier construction for each individual cluster.

REFERENCES

- [1] X. Liao and L. Carin, “Application of the theory of optimal experiments to adaptive electromagnetic-induction sensing of buried targets,” *IEEE Trans. Pattern Analysis Machine Intelligence*, vol. 26, pp. 961–972, Aug. 2004.
- [2] X. Liao, Y. Xue, and L. Carin, “Logistic regression with an auxiliary data source,” in *Proceedings of the 22nd International Conference Machine Learning (ICML)*, 2005.
- [3] M. Beal and Z. Ghahramani, “The variational Bayesian EM algorithm for incomplete data: Application to scoring graphical model structures,” *Bayesian Statistics*, vol. 7, pp. 453–464, 2003.
- [4] X. Zhu, “Semi-supervised learning with graphs,” Ph.D. dissertation, Carnegie Mellon University, 2005.
- [5] B. Krishnapuram, D. Williams, Y. Xue, A. Hartemink, L. Carin, and M. Figueiredo, “On semi-supervised classification,” in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2005.
- [6] Y. Zhang, L. Collins, H. Yu, C. Baum, and L. Carin, “Sensing of unexploded ordnance with magnetometer and induction data: theory and signal processing,” *IEEE Trans. Geoscience Remote Sensing*, vol. 41, pp. 1005–1015, May 2003.
- [7] W. Press, B. F. S. Teukolsky, and W. Vetterling, *Numerical Recipes in C : The Art of Scientific Computing, 2nd Edition*. Cambridge University Press, 1992.
- [8] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, “An introduction to variational methods for graphical models,” *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.
- [9] H. Attias, “A variational Bayesian framework for graphical models,” in *Advances in Neural Information Processing Systems (NIPS)*. MIT Press, 2000.
- [10] A. Dempster, N. Laird, and D. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of Royal Statistical Society B*, vol. 39, pp. 1–38, 1977.
- [11] M. Beal, “Variational algorithms for approximate Bayesian inference,” Ph.D. dissertation, Gatsby Computational Neuroscience Unit, University College London, 2003.

APPENDIX

A. Models for Spatially-Varying UXO Data

The spatially-varying EMI response (*i.e.*, measured data) of each anomaly is fit to two different models, a bivariate Gaussian model and a dipole-moment model. The dipole-moment model seeks to exploit the physics involved underlying EMI-based sensing, while the bivariate Gaussian model has been found to constitute a simple setting for extracting shape and size information about the EMI signature.

1) *Magnetic-Dipole Model*: A model for the frequency- and time-dependent EMI response of targets has been developed in [6]. Specifically, as a function of frequency, the magnetic-dipole moment \mathbf{m} of a target is represented as

$$\mathbf{m} = \mathbf{M}\mathbf{H}^{\text{inc}} \quad (10)$$

where \mathbf{H}^{inc} denotes the incident (excitation) magnetic field, and \mathbf{M} is the 3×3 dimensional magnetization tensor, relating the magnetic field to the magnetic-dipole moment. By taking the inverse Fourier transform of the frequency-domain EMI model, an analogous time-domain model can be obtained. This time-domain model is employed in this work because the sensor collects data in four time bins. For a UXO assumed to be rotationally symmetric with the axis of rotation along the z direction, the (time-dependent) magnetization tensor can be expressed as a diagonal matrix:

$$\mathbf{M} = \text{diag} \left[m_p \omega_p \exp \{-\omega_p t\}, m_p \omega_p \exp \{-\omega_p t\}, m_z \omega_z \exp \{-\omega_z t\} \right]. \quad (11)$$

In (11), the terms m_z and m_p correspond to the magnetic-dipole moments of the target, directed perpendicular to and along the target's axis of rotation, respectively; the terms ω_z and ω_p correspond to EMI “resonant” frequencies (decay constants), while t represents time.

If it is assumed that the EMI source responsible for the excitation magnetic field \mathbf{H}^{inc} can be represented — as seen from the target — as a magnetic dipole with moment \mathbf{m}_s , then [6]

$$\mathbf{H}^{\text{inc}} = \mathbf{r} \frac{1}{2\pi} \frac{\mathbf{m}_s \cdot \mathbf{r}}{|\mathbf{r}_{st}|^3}, \quad (12)$$

where \mathbf{r}_{st} is the vector directed from the source to the target center, with $\mathbf{r} = \frac{\mathbf{r}_{st}}{|\mathbf{r}_{st}|}$ the corresponding unit vector. Assuming sufficient proximity of the sensor's source and receiver

coils, the total (time-dependent) magnetic field observed at the sensor will be [6]

$$\mathbf{H}^{\text{rec}} \propto \frac{\mathbf{r}}{|\mathbf{r}_{st}|^6} \mathbf{r}^T \mathbf{U}^T \mathbf{M} \mathbf{U} \mathbf{r}, \quad (13)$$

where the proportionality constant depends on the strength of the dipole source \mathbf{m}_s and the characteristics of the receiver.

The 3×3 unitary rotation matrix \mathbf{U} rotates the fields from the coordinate system of the sensor to the coordinate system of the target, and \mathbf{U}^T transforms the dipole fields of the target (in the \mathbf{M} coordinate system) back to the coordinate system of the sensor. Explicitly, the target orientation, in terms of the angles of the target θ and ϕ with respect to the sensor coordinate system, is accounted for by

$$\mathbf{U} = \begin{bmatrix} \cos \phi & 0 & \sin \phi \\ 0 & 1 & 0 \\ -\sin \phi & 0 & \cos \phi \end{bmatrix} \begin{bmatrix} \cos \theta & \sin \theta & 0 \\ -\sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (14)$$

The EMI sensor employed in this work measures the z -component of the magnetic field as a function of position on the surface (z being normal to the air-soil interface). This measurement is subsequently fit to the model in (13) via a form of the Levenberg-Marquardt method [7]. Specifically, the parameters that the model inversion fits are the target position (x , y , and depth z), the target orientation (θ and ϕ), the magnetic dipole strengths (m_z and m_p), and the EMI resonant frequencies (ω_z and ω_p). We retain four parameters of the model — m_z , m_p , ω_z , and ω_p — as features for the classification stage.

To overcome the problem of local maxima, several (model-fitting) solutions are obtained, with each solution resulting from randomly initializing the parameters of the model. The final parameters of the model are taken to be those of the solution that minimize the mean-square error between the measured and model-fit data.

2) *Bivariate Gaussian Model:* The bivariate Gaussian model is given by

$$g(\mathbf{z}) = (2\pi)^{-1/2} |\Sigma|^{-1/2} \exp \left\{ -\frac{1}{2} (\mathbf{z} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{z} - \boldsymbol{\mu}) \right\} \quad (15)$$

where $\mathbf{z} = \begin{bmatrix} x & y \end{bmatrix}^T$ is the spatial position of the measured data, $\boldsymbol{\mu}$ and

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \quad (16)$$

are the free model parameters to be fit, and $g(\mathbf{z})$ is the spatially-dependent (scalar) data measurement of an anomaly. In general, the response of an anomaly will be comprised of measured data at many different spatial positions. This data is jointly fit using the model in (15). Subsequently, from this model fit, a single scalar feature that represents the size of the anomaly is computed as

$$f = \sigma_{11}\sigma_{22}|\boldsymbol{\Sigma}|^{-1/2}. \quad (17)$$

This size feature is used in later stages to cluster the anomalies.

B. Technical Details of Algorithms

1) *Active Learning*: In active learning, one selects the labels that would be most informative for building a classifier. Initially, one possesses no labeled data at a site. To utilize one particularly informative active-learning procedure, however, a classifier must already exist. Therefore, we first acquire labels for the anomalies associated with a set of basis vectors (the selection of which is discussed later). With this labeled data, a classifier can be constructed. Subsequently, we acquire additional labels under the purview of the active-learning process, with the goal of improving classifier quality.

Assume we possess a set of N training data points,

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N, \quad (18)$$

where \mathbf{x}_i is the i -th input vector, and y_i is the corresponding (scalar) output. The objective of kernel matching pursuits (KMP) is to obtain a linearly-weighted kernel-based regression function for which an input of \mathbf{x}_i would result in an output consistent with y_i . The desired regression function employing n basis functions is of the general form

$$f_n(\mathbf{x}) = \sum_{i=1}^n w_{n,i}K(\mathbf{x}, \mathbf{b}_i) + w_{n,0} = \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}_i), \quad (19)$$

where \mathbf{b}_i is the i -th basis function, and $\mathbf{w}_n = \begin{bmatrix} w_{n,0} & w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{bmatrix}^T$ are weights. In (19),

$$\boldsymbol{\phi}_n(\mathbf{x}_i) = \begin{bmatrix} 1 & K(\mathbf{x}_i, \mathbf{b}_1) & K(\mathbf{x}_i, \mathbf{b}_2) & \cdots & K(\mathbf{x}_i, \mathbf{b}_n) \end{bmatrix}^T, \quad (20)$$

wherein

$$K(\mathbf{x}_i, \mathbf{b}_j) = \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{b}_j)^2}{2\sigma^2} \right\} \quad (21)$$

is a kernel function that provides a measure of similarity between \mathbf{x}_i and \mathbf{b}_j .

The criterion used in active learning to choose which labels to acquire is based on an information-theoretic measure. As in [5], we define the matrix

$$\mathbf{H}_L = \sum_{i=1}^L p_i(1 - p_i)\phi_n(\mathbf{x}_i)\phi_n^T(\mathbf{x}_i), \quad (22)$$

where $p_i = p(y_i = 1|\mathbf{x}_i, \mathbf{w})$ is the classifier output giving the probability that the i -th of L labeled data points is a UXO, and where the classifier \mathbf{w} has been constructed using n basis vectors. If another label was acquired, the updated matrix would become

$$\mathbf{H}_{L+1} = \mathbf{H}_L + p_{L+1}(1 - p_{L+1})\phi_n(\mathbf{x}_{L+1})\phi_n^T(\mathbf{x}_{L+1}). \quad (23)$$

To determine which label should be acquired as the $(L + 1)$ -th label, we seek to maximize the gain in information, measured by the quantity

$$s(\mathbf{x}_{L+1}) = \frac{|\mathbf{H}_{L+1}|}{|\mathbf{H}_L|}. \quad (24)$$

Therefore, the $(L + 1)$ -th label to acquire corresponds to the unlabeled data point that would most increase the gain in information:

$$x_{L+1} = \arg \max_{\mathbf{x} \in \mathcal{D}_U} s(\mathbf{x}). \quad (25)$$

Intuitively, this formulation favors acquiring the labels of data points for which the current class predictions (*i.e.*, UXO or non-UXO) are most uncertain (*i.e.*, close to 0.5). It can also be shown [5] that in logistic regression frameworks, this formulation is equivalent to maximizing the mutual information between the newly selected label and the classifier weights.

It is also interesting to note that the true labels of the unlabeled data points are not required to compute the utility of acquiring the label. Because of this fact, one can form a list of excavations at once, and then plan the excavation path intelligently. That is, one need not determine the items to excavate one at a time.

Finally, the active-learning label-acquisition process terminates when the information gain is below a prescribed threshold. An example plot depicting the information gain for each unlabeled data point is shown in Figure 11; in this particular case, eleven or twelve new labels would be requested.

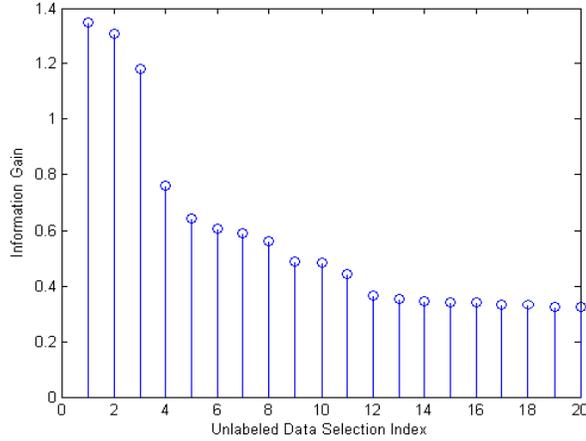


Fig. 11. Example information-gain plot for active learning.

2) *Variational Bayesian Gaussian Mixture Model (VB-GMM)*: Variational techniques are one family of methods that can be used to perform approximate inference [8]. A Gaussian mixture model (GMM) is a standard method to approximate a general distribution. In this work, a variational Bayesian Gaussian mixture model (VB-GMM) [9] is employed to cluster the anomalies at a UXO site into groups based on the size feature; the clustering is manifested by associating each feature vector with a particular GMM component, and therefore the mixture components correspond to clusters. Importantly, the VB-GMM formulation automatically determines an appropriate set of mixture (cluster) components, based on the observed data. As discussed in Section III, the results of this clustering step are important for analyzing the labeled data constituted via active learning, prior to design of the classifier.

Although in this work each anomaly is characterized by a scalar size feature, x_i , we shall describe the VB-GMM for the general (multivariate) case. Collectively, the data can be denoted $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$. Suppose that the distribution of this data is modeled as a mixture of K Gaussians:

$$p(\mathbf{x}_i|\Theta) = \sum_{k=1}^K p(\mathbf{x}_i|\gamma_i = k, \Theta)p(\gamma_i = k|\Theta) \quad (26)$$

$$= \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (27)$$

where $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$. With this formulation, the objective is to learn the set of

GMM parameters $\Theta = \{\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$. One standard approach to solving this problem is to learn the model parameters via an Expectation-Maximization (EM) algorithm [10]. In this work, we employ a variational Bayesian [3], [11] route. Whereas the result of the EM algorithm will be point estimates for the parameters, the result of the VB-EM algorithm will be *distributions* of the parameters. This fact allows the GMM learned via the VB-EM algorithm to be accurate even when faced with limited data.

Another significant drawback of the EM algorithm is that one must *a priori* choose the number of mixture components, K , that are represented by the data. By employing the VB-EM algorithm, one can determine the appropriate number of GMM components, K , in a principled manner. For instance, one can consider several different values for K , and for each value learn a GMM and compute the evidence [3] (which is an intermediate product of the VB-EM algorithm). The value for K that produces the maximum evidence is then chosen to be the true value of K . The variational formulation automatically penalizes overly complex models, whereas the standard EM algorithm will always favor increasingly complex models (*i.e.*, larger values of K). Consequently, the VB-GMM formulation determines the proper number of mixture components (clusters) based on the data.

With the use of the VB-EM approach motivated as above, we proceed to discuss the technical details of variational inference. Variational methods provide a lower bound on the log marginal likelihood, $\log p(\mathbf{X})$. In our GMM problem, \mathbf{X} are the observed variables, $\Phi = \{\gamma_i\}$ is the set of hidden variables, and $\Theta = \{\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ is the set of parameters. The log marginal likelihood of \mathbf{X} can be lower bounded by writing [3]

$$\begin{aligned} \log p(\mathbf{X}) &= \log \int p(\mathbf{X}, \Phi | \Theta) d\Phi d\Theta \\ &= \log \int q(\Phi, \Theta) \frac{p(\mathbf{X}, \Phi | \Theta)}{q(\Phi, \Theta)} d\Phi d\Theta \\ &\geq \int q(\Phi, \Theta) \log \frac{p(\mathbf{X}, \Phi | \Theta)}{q(\Phi, \Theta)} d\Phi d\Theta \end{aligned} \quad (28)$$

$$= \int q(\Phi) q(\Theta) \log \frac{p(\mathbf{X}, \Phi | \Theta)}{q(\Phi) q(\Theta)} d\Phi d\Theta \quad (29)$$

where (28) follows from Jensen's inequality, and (29) is the result of making the factorized approximation $q(\Phi, \Theta) \approx q(\Phi)q(\Theta)$. The variational Bayesian algorithm maximizes (29) with respect to the distributions $q(\Phi)$ and $q(\Theta)$. Since these two distributions are coupled, functional

derivatives with respect to each distribution are iteratively taken while the opposite distribution is held fixed. The resulting Variational Bayesian Expectation (VB-E) and Maximization (VB-M) steps are respectively

$$q(\Phi) \propto \exp \left\{ \int \log p(\mathbf{X}, \Phi | \Theta) q(\Theta) d\Theta \right\} \quad (30)$$

$$q(\Theta) \propto p(\Theta) \exp \left\{ \int \log p(\mathbf{X}, \Phi | \Theta) q(\Phi) d\Phi \right\}. \quad (31)$$

Recall that $p(\mathbf{X}, \Phi | \Theta) = p(\mathbf{X} | \Phi, \Theta) p(\Phi | \Theta) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{X}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. By choosing conjugate-exponential priors — a Dirichlet distribution on the mixing coefficients (π_k), normal distributions on the means ($\boldsymbol{\mu}_k$), and Wishart distributions on the precisions (inverse covariances, $\boldsymbol{\Sigma}_k^{-1}$) — the requisite integrals are tractable [9]. The VB-EM algorithm can then be employed, which provides the posterior distribution of the parameters of the GMM. Note that because of the use of conjugate priors, the priors and posteriors will be functionally identical (Dirichlet, Normal, and Wishart, for $\boldsymbol{\pi}$, $\boldsymbol{\mu}_k$, and $\boldsymbol{\Sigma}_k^{-1}$, respectively) but will have different parameter values. The equations for the updated form of the posterior parameters are given explicitly in [9], and therefore are not repeated here.

It should be emphasized that selecting the number of mixture components, K , in the variational formulation does not present limitations. In the variational formulation, K is the *maximum* number of mixture components to consider. When K is larger than the true number of mixture components represented by the data, the mixing proportions of the extraneous components will be very small ($\pi_k \rightarrow 0$). A simple solution to setting K is to choose a large enough value such that at least one mixing proportion is zero.

After the parameters of the GMM have been learned, one can easily assign each data point (here, an anomaly) to a single mixture component. In this manner, the anomalies are *clustered*, with each mixture component representing a unique cluster. It should also be noted that the low-dimension of the data, namely one in the present work, allows robust GMM parameters to be learned with very limited data. In Section II we discuss the use of such clustering to preprocess the labeled data determined via active learning, and explain why such an analysis is essential when applying active learning to field data.

3) *Kernel Matching Pursuits (KMP)*: For completeness, we here repeat a portion of the introductory material used in Appendix B.1, before explaining the specifics of the kernel matching pursuits classifier.

Assume we possess a set of N training data points,

$$\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N, \quad (32)$$

where \mathbf{x}_i is the i -th input vector, and y_i is the corresponding (scalar) output. The objective of kernel matching pursuits (KMP) is to obtain a linearly-weighted kernel-based regression function for which an input of \mathbf{x}_i would result in an output consistent with y_i . The desired regression function employing n basis functions is of the general form

$$f_n(\mathbf{x}) = \sum_{i=1}^n w_{n,i} K(\mathbf{x}, \mathbf{b}_i) + w_{n,0} = \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}), \quad (33)$$

where \mathbf{b}_i is the i -th basis function, and $\mathbf{w}_n = \begin{bmatrix} w_{n,0} & w_{n,1} & w_{n,2} & \cdots & w_{n,n} \end{bmatrix}^T$ are weights. In (33),

$$\boldsymbol{\phi}_n(\mathbf{x}_i) = \begin{bmatrix} 1 & K(\mathbf{x}_i, \mathbf{b}_1) & K(\mathbf{x}_i, \mathbf{b}_2) & \cdots & K(\mathbf{x}_i, \mathbf{b}_n) \end{bmatrix}^T, \quad (34)$$

wherein

$$K(\mathbf{x}_i, \mathbf{b}_j) = \exp \left\{ -\frac{(\mathbf{x}_i - \mathbf{b}_j)^2}{2\sigma^2} \right\} \quad (35)$$

is a kernel function that provides a measure of similarity between \mathbf{x}_i and \mathbf{b}_j .

The process of learning the desired regression function is composed of two parts. First, one selects a set of basis vectors (that are a subset of the training signatures in \mathcal{D}) that are statistically representative of the entire data set. After the set of basis vectors, \mathbf{B}_n , has been selected, the regression weights are adjusted in order to minimize a cost function, such as the error between the training data output, y_i , and the KMP model output, $f_n(\mathbf{x}_i)$. The aforementioned cost function is given by

$$e_n = \sum_{i=1}^N [y_i - f_n(\mathbf{x}_i)]^2 \quad (36)$$

$$= \sum_{i=1}^N [y_i - \mathbf{w}_n^T \boldsymbol{\phi}_n(\mathbf{x}_i)]^2. \quad (37)$$

The value of \mathbf{w}_n that minimizes the cost function (37) is

$$\mathbf{w}_n = \mathbf{M}_n^{-1} \left[\sum_{i=1}^N y_i \boldsymbol{\phi}_n(\mathbf{x}_i) \right], \quad (38)$$

where

$$\mathbf{M}_n = \sum_{i=1}^N \boldsymbol{\phi}_n(\mathbf{x}_i) \boldsymbol{\phi}_n^T(\mathbf{x}_i) \quad (39)$$

is the Fisher information matrix associated with the data and the basis vectors.

The basis vectors are selected in order to optimize an information-theoretic metric, measured in terms of the determinant of the Fisher information matrix [1]. Specifically, at each iteration, the data point (*i.e.*, feature vector) whose addition to the extant set of basis vectors would maximize the information gain is chosen.

The Fisher information matrix after including the $(n + 1)$ -th basis vector is

$$\mathbf{M}_{n+1} = \sum_{i=1}^N \begin{bmatrix} \phi_n(\mathbf{x}_i) \\ \phi_{n+1}(\mathbf{x}_i) \end{bmatrix} \begin{bmatrix} \phi_n^T(\mathbf{x}_i) & \phi_{n+1}^T(\mathbf{x}_i) \end{bmatrix} \quad (40)$$

$$= \begin{bmatrix} \mathbf{M}_n & \sum_{i=1}^N \phi_n(\mathbf{x}_i)\phi_{n+1}(\mathbf{x}_i) \\ \sum_{i=1}^N \phi_{n+1}(\mathbf{x}_i)\phi_n^T(\mathbf{x}_i) & \sum_{i=1}^N (\phi_{n+1}(\mathbf{x}_i))^2 \end{bmatrix}. \quad (41)$$

The gain in information accrued by including the $(n + 1)$ -th basis vector can be measured via the quantity

$$r(\mathbf{b}_{n+1}) = \frac{|\mathbf{M}_{n+1}|}{|\mathbf{M}_n|} = \left[\sum_{i=1}^N (\phi_{n+1}(\mathbf{x}_i))^2 \right] - \left[\sum_{i=1}^N \phi_{n+1}(\mathbf{x}_i)\phi_n^T(\mathbf{x}_i) \right] \mathbf{M}_n^{-1} \left[\sum_{i=1}^N \phi_n(\mathbf{x}_i)\phi_{n+1}(\mathbf{x}_i) \right]. \quad (42)$$

The $(n + 1)$ -th basis vector is thus chosen in a greedy fashion by selecting the data point that maximizes the information gain:

$$\mathbf{b}_{n+1} = \arg \max_{\mathbf{b} \in \mathcal{D}, \mathbf{b} \notin B_n} r(\mathbf{b}). \quad (43)$$

The basis vector selection process terminates when the information gain is below a prescribed threshold . With the set of basis vectors fixed, the classifier weights \mathbf{w}_n of the KMP algorithm are then given by (38). Passing the regression function output (33) through a sigmoid function then provides one with the probability of an unlabeled testing data point belonging to each class (UXO or non-UXO).