RANDOM SET FRAMEWORK FOR CONTEXT-BASED CLASSIFICATION

By

JEREMY BOLTON

A DISSERTATION PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

UNIVERSITY OF FLORIDA

2008

ACKNOWLEDGMENTS

I thank my mother, Lois Bolton, father, Wade Bolton, and sister, Chelsea Bolton for their relentless love and support.

I thank my advisor, Paul Gader, for his guidance and encouragement throughout my tenure at the University of Florida. I thank my committee Paul Gader, Joseph Wilson, Gerhard Ritter, Arunava Banerjee, and Clint Slatton for their insight and guidance which has steered my research and bettered resulting contributions.

I thank my lab mates for their support and am thankful for their ability to endure my shenanigans. I thank Alina Zare, Nathan VanderKraats, Nicholas Fisher, Xuping Zhang**,** Raazia Mazhar, Wen-Hsiung Lee and Seniha Esen Yuksel for their encouragement, suggestions and aid in my research.

I thank colleagues, Jim Keller, Hishem Frigui, and Dominic Ho, for their collaboration on a variety of research projects.

I thank William Clark of Army Research Office (ARO), Russell Harmon of ARO, Miranda Schatten of Night Vision and Electronic Sensors Directorate, and Michael Cathcart of Georgia Tech, for their support of my research.

TABLE OF CONTENTS

page

LIST OF TABLES

Table page

LIST OF FIGURES

Figure page

Abstract of Dissertation Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

RANDOM SET FRAMEWORK FOR CONTEXT-BASED CLASSIFICATION

By

Jeremy Bolton

December 2008

Chair: Paul Gader
Major: Computer Engineering

Pattern classification is a fundamental problem in intelligent systems design. Many different probabilistic, evidential, graphical, spatial-partitioning and heuristic models have been developed to automate classification. In some applications, there are unknown, overlooked, and disregarded factors that contribute to the data distribution, such as environmental conditions, which hinder classification.

Most approaches do not account for these conditions, or factors, that may be correlated with sets of data samples. However, unknown or ignored factors may severely change the data distribution making it difficult to use standard classification techniques. Even if these variable factors are known, there may be a large number of them. Enumerating these variable factors as parameters in clustering or classification models can lead to the *curse of high dimensionality* or sparse random variable densities. Some Bayesian approaches that integrate out unknown parameters can be extremely time consuming, may require *a priori* information, and are not suited for the problem at hand. Better methods for incorporating the uncertainty due to these factors are needed.

We propose a novel *context-based* approach for classification within a random set framework. The proposed model estimates the posterior probability of a class and context given

both a sample a set of samples, as opposed to the standard method of estimating the posterior given a sample. This conditioned posterior is then expressed in terms of priors, likelihood functions and probabilities involving both a sample and a set of samples. Particular attention is focused on the problem of estimating the likelihood of a set of samples given a context. This estimation problem is framed in a novel way using random sets. Three methods are proposed for performing the estimation: possibilistic, evidential, and probabilistic. These methods are compared and contrasted with each other and with existing approaches on both synthetic data and extensive hyperspectral data sets used for minefield detection algorithm development.

Results on synthetic data sets identify the pros and cons of the possibilistic, evidential and probabilistic approaches and existing approaches. Results on hyperspectral data sets in indicate that the proposed context-based classifiers perform better than some state-of-the-art, context-based and statistical approaches.

CHAPTER 1
INTRODUCTION

Problem Statement and Motivation

When collecting data, many known and unknown factors transform the observed data distribution. In many applications, sets of samples are collected at a given time, for example, remote sensing. In remotely sensed imagery, images are taken from a remote location such as a plane. These images are essentially sets of pixels, or samples, that are collected at the same time. In this instance, many of the unknown or unspecified factors may influence all of the samples in the image, or some subset thereof, similarly. That is, all of the samples in an image subset may undergo the same transformation induced by these factors.

Optical character recognition (OCR) is another application where factors may influence the results of classification. In OCR, if a classifier could identify a font or font size of a document, the problem of character recognition may be simplified. In this problem, the font or font size is a factor, or context, which may change the appearance of the sample, or the character.

Before we fully characterize the problem at hand, we state some assumptions and define a few terms which are necessary for the problem statement. We assume that similar samples collected in similar conditions or situations will undergo similar transformations. We define a *population* as a set of samples collected under the same conditions or situation. We define the idea of *context* as the surrounding conditions or situations in which data are collected. We define *contextual factors* as the unknown or unspecified factors that transform the data's appearance. Given these definitions, we can define a *contextual transformation* as a transformation that acts on sets of samples on a context-by-context basis. We attempt to estimate a population's context using the observed population's distribution.

In a probabilistic approach, context can be viewed as hidden random variables that are correlated with the observed samples. This view implies that the observed samples are dependent on these hidden variables.

In many standard models, classification accuracy suffers due to contextual factors. If these variables are ignored, many classification methods will suffer since the sample values may be severely altered by contextual transformations. On the other hand, if their values are specified and corresponding parameters are enumerated in a model, problems such as the curse of dimensionality or sparse probability distributions may hinder classification results.

Example 1.1 Contextual transformations: In this example, we illustrate that contextual factors are present in remotely sensed *hyperspectral imagery* (HSI) collected by *airborne hyperspectral imager* (AHI). In this data, each pixel in an image has a corresponding spectral vector, or *spectral signature*, with intensity values in the *long wave infrared* (LWIR), 7.8 um to 11.02 um. Each spectral signature is usually viewed as a plot of wavelength vs. intensity. Figure 1-1A illustrates multiple spectral signatures, or *spectra*, from a target class and a non-target class indicated by a solid line and a dashed line, respectively. Two consequences of contextual transformations can hinder classification. The first problem is the obvious change in sample appearance in varying contexts, which we refer to as a *non-disguising transformation*. An algorithm must know the appearance of a target sample for identification; therefore, if a target can potentially take on multiple appearances then a classifier must be aware of all potential appearances. The second problem occurs when samples from one class, in some context, are transformed to appear as samples from another class in another context, which we refer to as *disguising transformations*. We characterize these problems separately since their solutions require different approaches.

Solutions to non-disguising transformations require knowledge of the various target class appearances. An algorithm developer could simply add model constructs or parameters to account for varying appearances. For example, a developer could add densities to a mixture model to account for multiple appearances due to multiple transformations. However, this solution will not resolve the problem of disguising transformations since samples from different classes have the same appearance. In this situation, context estimation is used to identify relevant models that were constructed for similar contexts that our test population has been observed and thereby disregarding models or parameters constructed for irrelevant contexts.

Assume we want to classify the bolded spectral signature shown in Figure 1-1A. Classification is difficult since this spectral vector has the same appearance as some target and non-target spectra from various contexts. However, if we disregarded the spectra collected in a different context, classification becomes less complicated as illustrated in Figure 1-1B.

Example 1.2 Feature space transformation: Suppose we have images of scenes containing pixels with values in $\mathcal{R}^n$. For the sake of illustration, we assume n=2 and each image X, has a continuum of pixels. Each of these pixels corresponds to a measurement of some object in the real world. We would assume that the pixel's value would depend on the object it represents in the real world, but there are contextual factors that will influence the pixels' values.

In this example, there are five images containing pixels that represent two objects in the real world, 'x' and 'o'. Some of these images were taken in different contexts thus each is affected by different influencing factors. These contextual factors transform the data collected in distinct contexts, differently. These transformations may cause sets of samples to have different spatial distributions, or shapes, in a feature or sample space as shown in Figure 1-2A.

Assume the goal is to label some samples in $X1$, denoted by '*', using some labeled samples from the other images illustrated in Figure 1-2B. If we ignore the population information, the classification problem becomes more difficult as shown in Figure 1-2C. Instead, if we emphasize, to an algorithm, datasets which appear to have been collected in a similar context, the job of classification may be simplified, as shown in Figure 1-2D. A similar spatial distribution of sets may indicate that a similar transformation has acted on the populations and have therefore been collected in similar conditions. We propose that if this contextual information is gathered and utilized correctly, classification results should improve.

Proposed Solution

The problem of variable contextual factors is similar to some existing problems such as concept drift where the idea of a target class and/or its governing distribution may change with respect to time or some hidden context. In Example **1.2** , a solution would need to include a method for determining a similar distribution, or shape, relationship between populations. A more general solution would provide a method for modeling the shape of populations from a particular context.

Standard context-based classifiers suffer from a number of limitations. Most notably, they lack the ability to solve the problem of disguising transformations, as mentioned in Example **1.2**. Many classifiers attempt to estimate context, which we propose is best identified by analysis of an entire population, by inspecting a single sample. Many existing models also suffer from restrictions, inappropriate assumptions, and the lack of ability to handle all forms of concept drift. Most standard statistical methods make the *independently identically distributed* (i.i.d.)

assumption that limits their ability to capture any information found through the analysis of the set of samples.

The proposed solution uses a random set [1]-[7] model for population context estimation. A population's context is then considered when each sample of the population is classified. This model has the ability to estimate context by inspecting the distribution of a set of samples. Populations, after undergoing contextual transformations induced by contextual factors, are compared to contextual models—modeled using random sets—in attempts to identify the context in which they were collected. Specifically, the creation of the proposed context-based classifier consists of factors for context estimation and class estimation. The classification factor will estimate the class of each sample using class models, one for each context. The context estimation factor will identify the relevance of each model based on the estimated context of the test population and subsequently weight each model's contribution by contextual relevance. The identification of context allows for more informed class estimation emphasizing models relevant to the test population's context and ignoring the irrelevant models.

Note that the proposed model implicitly acquires context of a sample set without explicitly performing any estimation of the contextual factors. A subsequent benefit to this approach is that it avoids the curse of high dimensionality and sparse densities, which are potential pitfalls of methods that would directly account for these contextual factors.

The proposed random set model allows for evidential, probabilistic, and possibilistic approaches due to the inherent versatility of the random set. Furthermore, it also has the ability to avoid the aforementioned limitations and to handle all forms of concept drift. Existing standard and state-of-the-art methods are surveyed, analyzed, and compared to the proposed approach.

Results from experiments indicate that the proposed random set model improves classification results from existing methods in the face of hidden contexts.

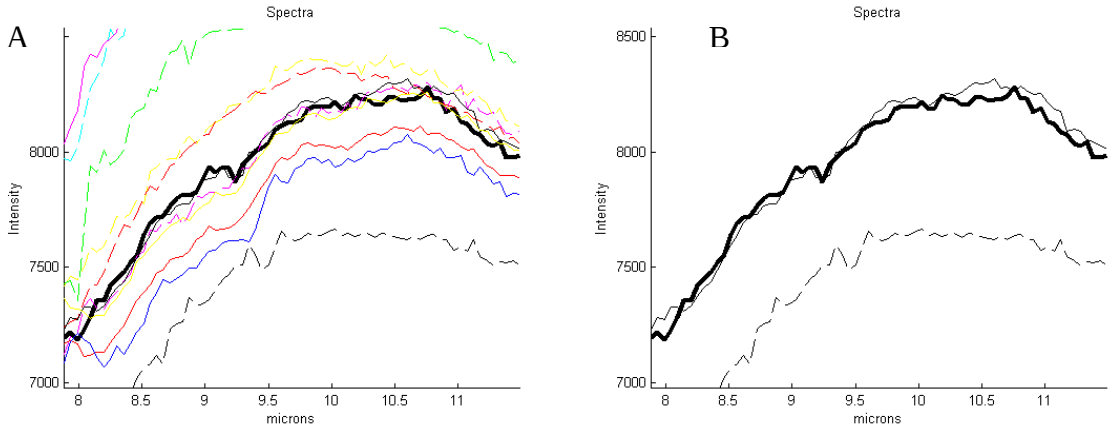

Figure 1-1. Spectral samples exhibiting contextual transformations. A) Spectra from target and non-target classes collected by AHI in multiple contexts. The target class is indicated by a solid line and a non-target class is indicated by a dashed line. B) An unlabeled sample shown in bold along with two labeled samples collected in the same context.

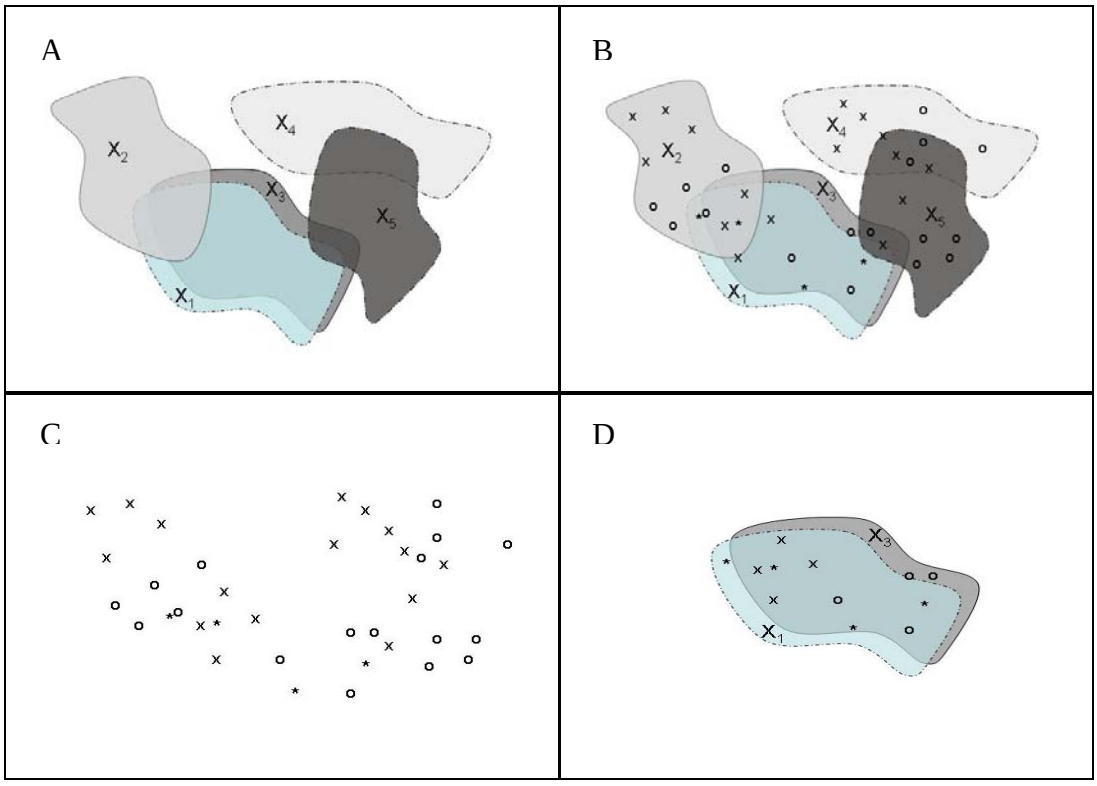

Figure 1-2. Illustration of contextual transformations in a feature space. A) Five images in some feature space that is a subset of $\mathcal{R}^2$. B) Labeled samples from each training image and unlabeled samples from the test image. C) All samples without contextual information. D) Using a similarly distributed training image to label the samples in the test image.

CHAPTER 2
LITERATURE REVIEW

The following is a review of current literature pertinent to problems and solutions arising from contextual factors. First, the problem of concept drift is detailed along with standard and state-of-the-art solutions [12]-[58]. Next, a brief review of context-based approaches with applications to hyperspectral imagery is given [59]-[67]. Next, a brief mathematical and statistical review is given to assist in the development of the proposed random set framework [1]-[11]. Standard statistical methods are reviewed and their potential uses for context-estimation are developed. Through the development we indicate that alternative methods may model the idea of context better than standard approaches. Next, the random set is defined and introduced as a method better suited for context estimation [1]-[7]. This is followed by a few examples of set similarity measures, which are reviewed to assist in set analysis [69]-[72]. Next, we review of some existing formulations and applications of random sets. Finally, we review some state-of-the-art, en masse, context-based approaches, which treat sets as unitary elements for context-estimation.

Concept Drift

The idea that samples of a class may change with respect to time is an area of recent research. We begin our discussion with a benchmark solution to this problem. One of the first algorithms developed to analyze and contend with this occurrence is STAGGER, which was developed by Schlimmer and Granger, and is based on a psychological and mathematical foundation [24]. STAGGER has 4 major steps: initialization, projection, evaluation, and refinement. In initialization, the description of a concept or class is constructed using a set of pairs consisting of logical statements, or characterizations, used to describe a class and corresponding weights used to weight the importance of each description. In this step, the

concept is specified. In projection, a Bayesian scheme is implemented to estimate the frequency of occurrences of the characterizations in subsequent samples. These probabilities are updated after the class of a new sample is determined. In this step, new samples are inspected to determine if frequency or weighting of each characterization is representative of the data. In evaluation, the effectiveness of each characterization is determined based on the number of correct and incorrect predictions for each characterization. In this step, the concept characterizations are evaluated to determine if there should be a change in these concept characterizations. In refinement, the characterizations and corresponding weights are modified based on their evaluations to improve their effectiveness as predictors.

The Problem of Concept Drift

STAGGER is one approach that contends with the change of concepts with respect to time or some hidden context. One of the more popular formulations of this problem, concept drift, has recently become an area of much research [18]-[57]. In concept drift, a concept may depend on some hidden context which is not given explicitly. Changes in the hidden context then induce changes in our target concept. This principle has been adopted by researchers in the machine learning community and has many applications in scientific research. Solutions to the problem should be able to adjust for concept drift, distinguish noise from concept drift and recognize and adjust for repeat concepts [18].

Concept drift can be divided into two categories: real and virtual. In real concept drift, the concept or idea of a target class may change. In virtual concept drift, the data distribution for a target class may change. The former is truly a *concept shift*—a change in concept—whereas the latter is simply a *sampling shift*—a change of data distribution due to some unknown context or variables. The idea of virtual concept drift is similar to our problem of hidden, population-

correlated variables, since this may lead to a change in data distribution due to some hidden context.

Concept drift can also be categorized as sudden or gradual. In sudden concept drift, the drift may be abrupt and substantial; whereas in gradual concept drift, the drift may be gradual and minimal. The problem at hand can be described as abrupt or sudden concept drift. The developed model allows for data to be collected at variable times and may not necessarily be a continuous flow of data with respect to time; in fact, the drift may be fairly substantial.

Concept Drift Solutions

There are three major approaches that are used to account for concept drift: instance selection, instance weighting and ensemble learning. In instance selection the goal is to select relevant samples from some training set for use in classifying test samples. A simple example of this approach would be windowing—using sliding windows or *k nearest neighbors* (kNN) [22], [23], [25]-[30]. Instance weighting involves weighting instances of a training set based on their relevance. Usually in instance weighting a learning algorithm is trained to appropriately weight these instances such as boosting [31]-[33], [39], [40]. In ensemble learning, a set of concept descriptions are maintained and some combination of these descriptions are used to predict current descriptions, as in STAGGER. This general approach could also be interpreted as some sort of model selection where the set concept descriptions are in fact models or algorithms whose results are to be combined based on each concept descriptions' relevance to a certain population [21], [24], [34]-[58].

In existing concept drift solutions, there are a number of restrictions, assumptions, and limitations that induce models that will not be able to account for all contextual transformations. Furthermore, almost all existing context-based solutions cannot solve the problem of disguising transformations as defined in Example 1.2. This drawback is due to the fact that context

estimation is performed by inspecting one sample, rather than the entire population. There are five major limitations or pitfalls exhibited by existing concept drift algorithms.

1. Estimates context based on a single sample (C.1)
2. Recognizes only some forms of concept drift (C.2)
3. Identifies context arbitrarily or with major assumptions (C.3)
4. Admits solutions that are not robust to outliers (C.4)
5. Assumes semi-supervised environment (C.5)

We emphasize property C.1 since this is a conceptual flaw implemented by many concept drift algorithms. This assumption presumes that the situation discussed in Example 1.2, disguising transformations, will not occur. Next, we survey standard and state-of-the-art approaches to concept drift. In the following, we parenthetically indicate where properties C.1 – C.5 are observed by the surveyed approaches. In almost all existing approaches, C.1 is present except when the approach is highly supervised and makes major assumptions for context identification.

Instance selection

In *full memory approaches*, all training samples are kept but a subset are selected to classify a given test sample. The process by which these samples are selected is the crux of instance selection approaches.

Widmer proposed the choice of a dynamic window size that is chosen based on time and classifier performance [30]. If the classifier is performing well, it is assumed that the concept has been constant for some time and a large window of samples are retained (C.2 and C.5). However, if performance decreases, it is assumed the concept is changing or has changed and the window size is shrunk (C.3 and C.4).

Klinkenberg *et al*. proposed an instance selection approach, where a variable sized window is kept over the *m* most recent training samples, assuming that the last *m* samples will be reflective of new test samples (C.2) [33], [34]. The selected window size minimizes the error of a support vector machine that is trained using the last *h* training samples. After the SVM is trained

using the last *h* samples, an upper bound on the error can be directly estimated from the SVM parameters [28], [7].

After these *m* SVMs have been trained on their last *h* samples, the training set with least error is selected. The window size is set to *h* as in Equation 2-1 and the corresponding training samples are used to classify the next test set.

$$h = \arg\min Err_{\xi\alpha}^{m}(h) \tag{2-1}$$

Here the SVM is used for an upper bound error estimate, and when its estimate increases, a change in context is assumed (C.3 and C.4).

Salganicoff proposed *Darling* which retains a selected sample until new samples are presented which occupy a similar subspace of the sample space [22]. This approach assumes context changes are directly related to the sequence of observance and that context is selected based on a single sample (C.1 and C.3).

Maloof *et al.* proposed an instance selection approach which is similar in ideology to instance weighting methods [26], [27]. In *partial-memory approaches*, each classification decision is made using some current characterization of a class and some subset of previously observed samples. The term partial-memory refers to the fact that only a subset of previously observed samples is retained to assist in classification and concept updating. Specifically in this method, the concept descriptions are updated using *selected* samples and misclassified samples [26]. Given a classifier *C*, a data set *D*, and a partial memory *P*, the update procedure consists of six major steps.

1. P={}
2. Classify *D* with *C*
3. Add misclassified samples to *P*
4. Retrain *C* using *P*
5. Select *appropriate P*
6. Repeat from step 2 when presented with new Data *D*'

Note that the classifier focuses on samples that it is misclassifying assumed to be due to concept drift (C.5 and C.4). An example of how to select an appropriate set *P* is to retain particular samples if they help form the decision boundary. One selection technique AQ-PM, which assumes a convex data set, identifies extreme points such as the points forming a covering hyperrectangle thus enclosing, or bounding, particular samples.

Instance weighting

Instance weighting approaches weight certain samples differently for the purposes of classification. A popular instance weighting scheme is boosting. A popular boosting algorithm is Adaptive Boosting, or *AdaBoost*, where misclassified samples are emphasized during parameter learning stage in a statistical manner [30]-[33]. The error term is calculated as follows:

$$\varepsilon_t = \sum_{i=1}^{n} D_t(i)[y_i \neq C_t(x_i)]. \tag{2-2}$$

In Equation 2-2, *t* is the learning iteration, x_i is sample *i*, $y_i \in \{-1,1\}$ is the class for x_i, C_t is the classifier at iteration *t*, $D_t(i)$ is the weight for sample x_i at iteration *t*, and ε_t is the average misclassification at iteration *t.* If the classifier misclassifies some samples, assumedly due to concept drift (C.3and C.5), the misclassified samples are emphasized (C.4) in the error term using the weight update formula.

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-\alpha y C_t(x_i)\right)}{Z_{t+1}} \tag{2-3}$$

This update increases the weights of misclassified samples to coerce the learning of the new concept in later iterations. Note this is similar to increasing the prior of x_i, in the statistical sense. Note if the boosting is done offline, just during training, this approach no longer exhibits property C.5, and maybe not C.4; however, it will exhibit property C.1.

Dura, Lui, Zhang, and Carin proposed neighborhood-based classifiers where a test sample's neighborhood is used for classification [35]-[38]. This approach uses and *active learning* framework which attempts to extract information from some dataset and extend it to another sample under test (C.1). Classification is performed as shown in Equation 2-4.

$$p(y_i \mid N(\mathbf{x}_i), \boldsymbol{\theta}) = \sum_{j=1}^{n} b_{ij}\, p(y_i \mid \mathbf{x}_j, \boldsymbol{\theta}),\ \ p(y_i \mid \mathbf{x}_j, \boldsymbol{\theta}) = \frac{1}{1+\exp\left(-y_i \boldsymbol{\theta}^T \mathbf{x}_j\right)}. \tag{2-4}$$

In Equation 2-4, $y_i \in \{-1,1\}$ is a class label, $\mathbf{x}_i$ is a test sample, $\mathbf{x}_j$'s are retained samples that are in the neighborhood $N(\mathbf{x}_i)$, b_{ij}'s are the weights for each neighbor, and $\boldsymbol{\theta}$ is a parameter vector. The construction of b_{ij}, the weight, and $N(\mathbf{x}_i)$, the neighborhood, are the crux of this algorithm. A few suggestions are shown in Equations 2-5 and 2-6.

$$N_t(\mathbf{x}_i) = \{(\mathbf{x}_j, b_{ij}) : b_{ij} > 0, \mathbf{x}_j \in X\} \tag{2-5}$$

$$\text{where } b_{ij}^t = \frac{\exp(-.5\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|}{\sigma_i^2})}{\sum_{k=1}^{n} \exp(-.5\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|}{\sigma_i^2})}. \tag{2-6}$$

In Equations 2-5 and 2-6, b_{ij} is the transition probability from $\mathbf{x}_i$ to $\mathbf{x}_j$ in less than *t* steps in Markov random walks [36], [37].

In some of their other proposed methods, an information theoretic approach is taken to construct $N(\mathbf{x}_i)$ based on maximizing the determinant of the Fisher information matrix [35], [38]. Note this approach also exhibits property C.1 since each sample is classified using itself and training data, not its population.

Note since the parameter $\boldsymbol{\theta}$ doesn't vary, we assume there is only one concept descriptor, which is why we consider this an instance weighting approach. We note that approach could also be implemented using an ensemble learning approach.

Ensemble learning

In ensemble learning an ensemble of concept descriptions, such as classifiers, are maintained and used in harmony for classification. A popular approach, ensemble integration, employs a weighted scheme to determine the relevance of each classifiers output given a sample [41].

$$C(x_i) = \sum_{j=1}^{J} w_{ij} C_j(x_i) \tag{2-7}$$

Here the construction of the weight w_{ij} is done to emphasize classifiers of greater contextual relevance. Equation 2-7 can be implemented in many ways such as static voting/weighting or dynamic voting/weighting [39]-[58]. In ensemble approaches, the crux of the problem is deciding how to weight each context-based model.

The popular *bagging* approach constructs *N* classifiers where each are trained using *N* corresponding training sets [43]. The training sets are constructed by randomly sampling the entire training set with replacement. Each of the sampled training sets contains *m* samples where *m* is less than the number of total training samples. The classifiers, which act on individual samples, are then combined using voting and averaging techniques (C.1 and C.3).

The random forest model is a new approach using dynamic classifier integration [44], [45], [47]. This model attempts to minimize correlation between the individual classifiers while maintaining accuracy [43], [44]. Random subspaces and/or subsets of samples are chosen and a classifier, or tree, is trained using the corresponding samples (C.3). This is repeated *N* times to create a forest of *N* trees. Most of the time the classifiers are simply partitionings of the space resulting in boolean classification. Given a test sample, classification is determined by weighting each tree's confidence using the confidences of neighboring samples in the feature space (C.1 or

C.3 and C.5 depending on implementation). The weight w_i for tree *i* is assigned using Equation 2-8.

$$w_i(x) = \frac{\sum_{j=1}^{k} 1_{OOB_i}(x_j)\phi(x, x_j)mr_i(x_j)}{\sum_{j=1}^{k} 1_{OOB_i}(x_j)\phi(x, x_j)} \tag{2-8}$$

In Equation 2-8, $mr_i(x_j) \in \{-1,1\}$ indicates whether classifier *i* has correctly classified sample *j*, ϕ is a weighting function based on distance, *k* is the size of the neighborhood, and 1_{OOB_i} is the indicator function, which indicates whether its argument is an *out-of-bag* (OOB) sample—a sample not used to train classifier *i*.

The use of OOB samples allows for unbiased estimates. We note that given some assumptions, the random forest approach is shown to perform at least as well as boosting and bagging [44].

Tsymbal *et al.* proposed an ensemble approach that maintains a set of models optimized over different time periods to handle local concept drift (C.2) [21], [39]. The models predictions are then combined, in a sense integrating over classifiers. The selection of classifier predictions is done based on a local classification error estimate performed after initial training. During testing, *k* nearest neighbors of each test sample are used to predict the local classification errors of each classifier (C.1). Using these estimated errors, each classifier's predictions are weighted and the total prediction is calculated using integration.

Kuncheva and Santana *et al.* developed an ensemble approach where contexts or training sets are constructed by clustering the training data [48], [49]. Then for each cluster, *N* classifiers are ranked such that each has a ranking in each cluster—set of samples. The weights for

combination are proportional to the classifiers correct classification. A test sample is then classified using the *k* best classifiers from the sample subspace in which it resides (C.1).

Frigui *et al.* used fuzzy clustering methods to partition a feature space into assumed contexts [52]. During classification, the models representing a context in which a test sample lies are used for classification where the classifiers are weighted by the corresponding fuzzy memberships of the test sample to the fuzzy cluster (C.1).

Harries *et al.* proposed an algorithm to learn hidden contexts called *Splice* [57], [58]. In this algorithm, a continuous dataset is partitioned, heuristically, into time intervals which supposedly represent partial contexts. Classifiers are then trained and ranked on each interval. The intervals, and classifiers, are then clustered similarly to an agglomerative clustering algorithm. If a classifier performs well on multiple contexts, the corresponding contexts and classifiers are merged and the classifiers are re-ranked based classification results. The weights are then selected similarly to the approaches proposed by Kuncheva and Santana *et al.* (C.1) [48], [49].

Santos *et al.* proposed a subsetting algorithm that randomly creates subsets of the training data (C.3) [50]. A classifier is trained on each subset, assumed to be indicative of a context, and a genetic algorithm selection scheme is used to select the best fit classifiers, where fitness is based on error rate, cardinality, and diversity. Context models are then weighted based on which subset a test sample resides (C.1).

Qi and Picard proposed a context-sensitive Bayesian learning algorithm that models each training set as a component in a mixture of Gaussians [55]. In this model each training set, or context, has a corresponding linear classifier.

$$p(y \mid x, D) = \sum_{i \in I} p(y \mid x, D_i)\, p(D_i \mid x, D) \tag{2-9}$$

In Equation 2-9, y is the class label for sample x using training dataset $D_i \in D = \{D_1,...,D_I\}$. The term $p(y \mid x, D_i)$ is estimated using the expectation propagation method [56]. Note the data set weights are chosen based solely on the sample x and not the sample and its population (C.1). Also, note that each D_i are training sets and not necessarily the population of sample x.

In the proposed random set model for context based classification, test sets are used to estimate context which alleviates property C.1, and furthermore does not induce properties C.2-C.5.

Applications to Hyperspectral Imagery

In the experiments, the proposed methods are tested using a hyperspectral dataset with apparent contextual factors. For this reason, we briefly discuss current, state-of-the-art methods used to contend with contextual factors in hyperspectral imagery. We note that some methods take different approaches or assume a different testing environment.

There are two major approaches for solutions to contextual transformations in hyperspectral data classification. The first approach relies on physical modeling using environmental information. The other uses statistical and/or mathematical methods to identify or mitigate the effects of contextual transformations. Next, we list some popular existing approaches which have shown to be successful in some testing situations.

There has been much research that uses the physical modeling of the environmental factors on measured data. Here, classifiers may use the output of physical models, for example MODTRAN, which generate the appearance of target spectra in certain environments [59], [60]. For example, the hybrid detectors developed by Broadwater use target spectra that are estimated using MODTRAN, which is given environmental information about the scene [61]. This

approach, and many like it, are shown to be very successful when environmental conditions are available.

Healy *et al.* proposed to use MODTRAN to produce spectra of various materials in various environmental conditions [62]. A vector subspace for each material is then defined by selecting an orthonormal basis for the material subspace. Confidence is then assigned to test spectra based on their distance to this subspace. This approach provides a robust and intuitive solution; however, this classification method will suffer in the presence of disguising transformations.

Kuan *et al.* proposed a projection matrix, rooted in a physics-based linear reflectance model, which in effect normalizes environmental conditions between two images [63]. This approach has shown to be successful at identifying regions of images and detecting change in co-registered imagery. This approach can learn a transformation of a set of samples; however, this approach requires a fairly large number of test sample labels be known for the construction of the transformation matrix.

Fuehrer *et al.* proposed the use of atmospheric sampling where a sample of some material is projected into some feature space based on atmospheric conditions in which it was observed [64]. Samples in this feature space may then be used to assist, using locality analysis, in identifying material and atmosphere when presented with a test image. This method has shown to be successful at classification and modeling; however, it cannot account for disguising transformations.

In these approaches, environmental conditions of a scene are assumed to be known *a priori*, or some ground truth is assumed to be known *a priori*, which may not be the case. In these other cases, different approaches must be taken.

The other tactic of existing methods uses various statistical and mathematical approaches to account for contextual transformations. Some selection, ensemble, and context-based methods attempt to identify models relevant to a test sample through context estimation. Some active learning approaches attempt to transfer knowledge to test samples.

Mayer *et al.* propose the whitening / dewhitening transformation. In this approach, transformation matrices are constructed to whiten and dewhiten spectra from an image [65]. In this approach, the whitening and dewhitening matrices are constructed to whiten the effects of environmental conditions. However, this approach requires a semi-supervised testing environment to construct the projection matrix. It also assumes that whitening of spectra will reduce or eliminate the effects of contextual factors. This assumption implies that the contextual transformation is simply a linear transformation based on a population's statistical properties, such as the mean and covariance. Mayer proposes the matched filter described in Equation 2-10.

$$MF_k = \left(x_{t,k} - \bar{x}_t\right)^T R_{22}^{-1}\left(L_t^{Transform} - \bar{x}_t\right), \quad \text{where } L_t^{Transform} - \bar{x}_t = R_{tt}^{1/2} R_{11}^{-1/2}(L_1 - \bar{x}_1). \tag{2-10}$$

In Equation 2-10, $x_{t,k}$ is a test sample, $\bar{x}_1$ is the mean of clutter samples from labeled image 1, $\bar{x}_t$ is the mean clutter estimate from the test image, L_1 is the target estimate for labeled image 1, and R_{11}, R_{tt} are the clutter covariance matrices for image 1 and the test image, respectively.

Rajan *et al.* propose an active learning approach where a classifier, or learner, attempts to acquire knowledge from a teacher about new data points that may be from an unknown distribution [66]. In this so called *KL-max* approach, the new data points and corresponding labels are chosen to maximize the KL divergence between the learned distributions and the learned distributions including the new data point and corresponding label. The labels, which are distributions, are then updated using the new data point and label. This approach could be used

for context estimation where various labels from existing classifiers are chosen based on the KL divergence; however, it estimates these labels sample-by-sample.

Many of the aforementioned existing methods either operate in different testing conditions, such as semi-supervised classification or environmental conditions are known *a priori*, or they cannot account for disguising transformations.

Probability Introduction

We now provide a brief mathematical and probabilistic review of the concepts that will be used in the proposed model. Due to the complex formulation of random sets, our review starts with the building blocks of probability and measure theory. The main purpose of the following review is the introduction of notation. For a rigorous mathematical development, see the literature [1]-[7].

Informally, a random variable is a mapping from a probability space to a measurable space. The probability space consists of a domain, family of subsets of the domain, and a governing probability distribution. To formally define random variables, we need to introduce concepts from topology and measure theory.

Topology

Definition 2.1 Topology: A topology T on a set X is a collection of subsets of X that satisfy
1. $\phi, X \in T$,
2. T is closed under finite unions and arbitrary intersections.

Such a pair, (X,T), is referred to as a topological space [10].

The set X is subsequently referred to as a topological space. Topologies are generally described by construction. Usually, a topology is said to be generated from some basis or sub-basis B.

Definition 2.2 Basis for a topology: A basis for a topology T on X is a collection $\mathcal{B}$ of subsets of X such that

1. For all $x \in X$ there exists a $B \in \mathcal{B}$ such that $x \in B$.
2. If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$ then there exists a B_3 such that $x \in B_3$ and $B_3 \subseteq B_1 \cap B_2$ [10].

Definition 2.3 Subbasis for a topology: A subbasis for a topology on *X* is a collection of subsets of *X* whose union is *X*. The topology generated by a subbasis *S* is the collection *T* of all unions and finite intersections of the elements of *S* [10].

The constituent sets of a topology are the focus of this review. Therefore, we fully detail them and the idea of measurability.

Definition 2.4 Open set: Given a topological space (*X*,*T*), all sets $G \in T$ are called open sets [10].

Definition 2.5 Closed set: The complement of an open set is a closed set [10].

A major misconception is that sets are either closed or open; however, this is not the case. In fact sets in a topology can be open, closed, neither, or both. For instance in the standard topology on $\mathcal{R}$, the interval [0,1) is neither open nor closed [10]. We emphasize that this is greatly dependent on how the topology is generated. There are topologies that do not share the intuitive characteristics of the standard topology on $\mathcal{R}$.

We next define some attributes of a topological space, which help characterize important concepts. Many of these attributes such as compactness are assumed when dealing with sets, but in the following, they are formally defined for clarity.

Definition 2.6 Cover: A collection of subsets of a space *X* is said to cover *X* if the union of its elements is *X*. Furthermore, an open cover of *X* is a cover whose elements are open sets [10].

Definition 2.7 Connectedness: A topological space (*X*,*T*) is connected if there does not exist a pair of disjoint, non-empty, open subsets *U* and *V* of *X* whose union is *X* [10].

Definition 2.8 Compactness: A space is compact if every open covering of *X* contains a finite subcollection that also covers *X* [10].

Probability Space

Next, we define necessary constructs for a probability space. We then define a standard random variable which will aid in the development of the random set.

Definition 2.9 σ-Algebra: If X is a set, then a σ-algebra $\sigma(X)$ on X is a collection of subsets of X that satisfy

1. $X \in \sigma(X)$
2. $A \in \sigma(X) \Rightarrow A^C \in \sigma(X)$
3. If $\{A_n\}_{n=1}^{\infty}$ is a sequence of elements of $\sigma(X)$, then $\bigcup_{n=1}^{\infty} A_n$ is $\sigma(X)$. Furthermore, $\sigma(X)$ is closed under countable intersections [9].

Note that if $\{A_n\}$ is a finite or countably infinite collection of elements of $\sigma(X)$, then $\left(\bigcup A_n\right)^c = \bigcap A_n \in \sigma(X)$ thus a σ-algebra is also closed under countable intersections. Hence, σ-algebras are topologies since the requirements for σ-algebras subsume the requirements of topologies. Note that σ-algebras also require closure under complementation, which is not a requirement of a topology. This closure under complementation allows for an intuitive application to probabilistic analysis. A σ-algebra is a type of topology useful in the field of probability and measure theory. In fact, most probability spaces are defined using Borel σ-algebras.

Definition 2.10 Borel σ-algebra: The Borel σ-algebra on a topological space X, written $\mathcal{B}(X)$, is the smallest σ-algebra that contains the family of all open sets in X.

Elements of a Borel σ-algebra are called *Borel Sets*.

Measure

Before we introduce random variables, we explain the idea of measurability. Although the general idea of measure is fairly complex, we give a simple overview.

Definition 2.11 Measure: A measure on $\sigma(X)$ is a function $\mu : \sigma(X) \to [0, \infty)$ satisfying

1. $\mu(\phi) = 0$
2. $A \cap B = \phi \Rightarrow \mu(\{A \cup B\}) = \mu(A) + \mu(B), \forall A, B \in \sigma(X)$, if finite or

$A_n \in \sigma(X), \forall j \neq k \;\; A_j \cap A_k = \phi \Rightarrow \mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$ if infinite [9].

The elements of $\sigma(X)$ are called measurable sets [9].

Some measures have added constraints such as the probability measure.

Definition 2.12 Probability measure: A probability measure is a measure $P:\sigma(X)\to[0,1]$ with the added constraint $P(X)=1$.

We have now properly defined the probability measure which is one of three elements necessary for a probability space. The other two elements are the domain and a corresponding σ-algebra.

Definition 2.13 Measure space: A measure space is a triple $(X,\sigma(X),\mu)$ where the pair $(X,\sigma(X))$ is referred to as the measurable space, *X* is a topological space, $\sigma(X)$ is a σ-algebra on *X*, μ is a measure on $\sigma(X)$ [9].

Definition 2.14 Probability space: A probability space is a triple $(\Omega,\sigma(\Omega),P)$, where Ω is a topological space, $\sigma(\Omega)$ is a σ-algebra on Ω, and *P* is a probability measure on $\sigma(\Omega)$ [9].

Definition 2.15 Measurable function: A function $f:\sigma(X)\to\mathcal{A}$ is measurable if for any interval $A\subseteq\mathcal{A}$, $f^{-1}(A)\in\sigma(X)$ [9].

A random variable is a measurable mapping from some probability space into a measurable space.

Standard Random Variables

Random variables are the basis of statistical modeling and analysis. The use of statistical modeling and analysis is abundant in the pattern recognition and machine learning community. These tools, along with others, allow researchers to model systems and automate *intelligent* decision making.

Now that we have defined all the necessary structures, we are able to define the random variable.

Definition 2.16. Random variable: Given a probability space $(\Omega,\sigma(\Omega),P)$ and some measurable space $(X,\sigma(X))$ for some positive integer *d*, a random variable, *R*, is a measurable mapping from a probability space to a measurable space such that $\forall \mathrm{Y}\in\sigma(X), R^{-1}(\mathrm{Y})\in\sigma(\Omega)$, if the random variable is defined on the entire space [9].

We note here that in applications, many ignore this initial mapping from the probability space to the measurable space. This mapping is necessary for formal definitions; however, it is

not necessary for most applications and the cumbersome notation is disregarded. Hereafter, we may disregard this initial mapping unless its recognition is required.

Standard Statistical Approaches for Context Estimation

There are a few issues that will arise if standard statistical techniques are used for context estimation. Next, we detail some of these potential pitfalls.

In standard approaches, the probability or likelihood of multiple occurrences are calculated using a joint distribution

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid C) \tag{2-11}$$

where $x_1, x_2, ..., x_n$ are *n* observations and *C* is some context. A few issues that arise from this approach are as follows:

1. Estimation of the joint likelihood function may be complicated by sparsity (J.1)
2. Estimation requires the matching of observations to random variables (J.2)
3. Likelihood calculation is highly dependent on number of observations (J.3)

Issue J.1 will occur when there are a large number of random variables compared to number of observations. Issue J.2 occurs since there is a distinction made between the observations. If X_i is different from X_j, then each observation will have to be paired with a random variable. This presents a problem of matching each observation to a random variable which also results in issue J.3.

Standard random variables are used to model the outcomes of single events or trials. In some approaches, a set of observations is modeled using a standard random variable where the set of observations is interpreted as a sequence of trials from the same experiment. This approach is similar to a common assumption for simplified joint estimation, the i.i.d. assumption.

$$P(x_1, x_2, ..., x_n \mid C) = P(x_1 \mid C)P(x_2 \mid C)...P(x_n \mid C) \tag{2-12}$$

This assumption presumes that observations *x* can be fully described by one random variable. However, this simplification results in a two additional issues:

1. Estimate of the joint likelihood is not robust to outliers due to the product of sample likelihoods (J.4)
2. Contextual information concerning joint observation is reduced to a product of sample likelihoods (J.5)

Note that even with the i.i.d. assumption, issue J.3 is still present. For example, as the number of observations occurs, the likelihood of some context must decrease, which is an unintuitive result for modeling context. This result is intuitive if we are modeling a sequence of experiments. Issue J.4 occurs since we have turned joint estimation into a product of singleton likelihoods.

Random Sets

One type of random variable—the random set—has not been researched as extensively as the standard random variable in the intelligent systems community. We consider only random subsets of $\mathcal{R}^d$ in the following. First, the formal definition of the random set and some associated constructs are presented. Next, a brief inspection and discussion of the random set is presented including its relationship to belief and possibility theory. Finally, the shortcomings of standard point process models for context estimation are discussed, which provides motivation of the proposed implementations.

General Case: Random Closed Set

Assume that $\mathcal{E} \subseteq \mathcal{R}^d$ is a topological space. We will denote the family of closed subsets of $\mathcal{E}$ as $\Im$. We can define a measurable space $(\Im, \sigma(\Im))$ associated with some probability space $(\Omega, \sigma(\Omega), P)$, where all $\Im$-valued elements will be referred to as closed sets. Informally, a random set is a measurable mapping from the aforementioned probability space to the measurable space.

Note that the construction of an intuitive σ-algebra for closed set values is not as clear as the construction for real number values. For example, a measurable interval for a random variable may be [-1, 4]. This interval, or set, is constructed by accumulating all the numbers *greater than or equal to* -1 and *less than or equal to* 4. However, relationships such as *greater than* or *less than* do not linearly order sets. One σ-algebra that is used with random sets is constructed by the Hit-Miss or Fell topology, such that any observed set $X \in \Im$ either intersects, hits, or does not intersect, misses, some $K \in \mathcal{K}$, where $\mathcal{K}$ is the family of compact sets. The families of sets that are used as basis elements to generate the Fell topology are $\Im^K = \{F \in \Im : F \cap K = \phi, K \in \mathcal{K}\} \in \sigma(\Im)$ and $\Im_G = \{F \in \Im : F \cap G \neq \phi, G \in \mathcal{G}\} \in \sigma(\Im)$. The Fell topology is a standard topology on $\Im$.

Definition 2.17 Fell topology: The Fell topology is a topology $(\Im, T)$ where *T* has subbasis which consists of $\Im_G$ and $\Im^K$.

Note that the Borel σ-algebra generated by the Fell Topology on $\Im$ coincides with the σ-algebra generated by $\Im_K$ [1]. We can now formally define the random closed set.

Definition 2.18 Random closed set measurable with respect to the Fell topology: Let $\Im$, be a collection of all closed sets from a topological space and let $\mathcal{B}(\Im)$ denote the σ-algebra generated by $\Im_K$. Given a measurable space $(\Im, \mathcal{B}(\Im))$ associated with some probability space $(\Omega, \sigma(\Omega), P)$,a measurable mapping $\Xi : \Omega \to \Im$ is called a random closed set measurable with respect to the Fell Topology if $\mathcal{B}(\Im)$ [1].

Random Set Discussion

The random set is governed by its distribution $P(\Im_K) = P\{\Xi \in \Im_K\}, \Im_K \in \mathcal{B}(\Im)$. Since $\mathcal{B}(\Im)$ is generated by $\Im_K$, it seems reasonable to determine the measure, or probability, of some set *K* using $\Im_K$ where $P\{\Xi \in \Im_K\} = P\{\Xi \cap K \neq \phi\}$ is a well defined measure. In fact, since these sets $\Im_K$ for each *K* compose our Borel σ-algebra, our probability distribution is defined on these sets with corresponding values being the probability of an observed Ξ will intersect *K*.

Note that the sets in $\Im_K$ just have to have a non-empty intersection with some set value *K*. In effect, the calculation of likelihood of a random set value *K* can be viewed as calculating the measure of the sets that contain at least one similar component as the set *K*.

Definition 2.19 Capacity functional: The real-valued function, T_Ξ ,associated with Ξ, $T_\Xi(K) = P(\Xi \in F_K) = P\{\Xi \cap K \neq \phi\}, K \in \mathcal{K}$ is called the capacity functional if the following requirements are satisfied [1]:

1. $T_\Xi(\phi) = 0$
2. $0 \leq T_\Xi(K) \leq 1, K \in \mathcal{K}$
3. $K_n \downarrow K \Rightarrow T_\Xi(K_n) \downarrow T_\Xi(K)$ (upper semi-continuous)
4. $\Delta_{K_n} \ldots \Delta_{K_1} T_\Xi(K) \leq 0, \ \forall n \geq 1, \ K, K_1, ..., K_n \in \mathcal{K}$ (completely alternating/ completely $\cup$ – alternating)

where $\Delta_{K_n} \ldots \Delta_{K_1} T_\Xi(K) = -P\{\Xi \cap K = \phi, \Xi \cap K_i, i = 1, ..., n\}$.

For an extensive explanation, the reader is directed to literature [1]-[6].

The capacity functional can be viewed as an optimistic estimate of the probability of a random set. In fact, it can be shown that this measure is an upper bound the family of probability measures $\mathbf{P}_\Xi$ associated with random set Ξ, that is $T_\Xi(K) = \sup\{P(K) : P \in \mathbf{P}_\Xi\}$ [1]. This also means that the capacity functional is an upper probability. It can be shown that $T_\Xi(K)$ dominates $P(K)$, $\forall P \in \mathbf{P}_\Xi$, which means $T_\Xi(K) \geq P(K), \forall K \in \mathcal{K}, \ \forall P \in \mathbf{P}_\Xi$ [1].

To uncover other functionals associated with the random set, we dissect the set $\Im_K$ into three disjoint sets.

$$\Im_K = \{F \in \Im : F \subset K\} \cup \{F \in \Im : K \subset F\} \cup \{F \in \Im : F \cap K \neq \phi, F \not\subset K, K \not\subset F\} \qquad (2\text{-}13)$$

Since the constituent sets in Equation 2-13 are disjoint, we can divide the capacity functional into these following terms:

$$\begin{aligned} &P\{\Xi \cap K \neq \phi\} = P\{K \subset \Xi\} + P\{\Xi \subset K\} + P\{\Xi_i \cap K \neq \phi, \Xi \not\subset K, K \not\subset \Xi\} \\ &= I_\Xi(K) + C_\Xi(K) + H_\Xi(K). \end{aligned} \qquad (2\text{-}14)$$

Note that T_{Ξ} is not additive with respect to *K*, but rather partitions of $\Im_K$. For example, if $K = K_1 \cup K_2, K_1 \cap K_2 = \phi$, then $P\{\Xi \cap K_1 \neq \phi\} + P\{\Xi \cap K_2 \neq \phi\} \neq P\{\Xi \cap K \neq \phi\}$ may be possible. This is true since it may be the case that $\exists K_3$ such that $K_3 \in \Im_{K_1} \cap \Im_{K_2}$; and by definition, $K_1 \cap K_2 = \phi$ does not imply $\Im_{K_1} \cap \Im_{K_2} \neq \phi$. In fact, T_{Ξ} is a subadditive fuzzy measure on $\Im$,

$$P\{\Xi \cap K \neq \phi\} \leq P\{\Xi \cap K_1 \neq \phi\} + P\{\Xi \cap K_2 \neq \phi\}. \tag{2-15}$$

We now define the functionals developed in Equation 2-14.

Definition 2.20 Inclusion functional: The inclusion functional calculates the measure of the sets in which *K* is included—all the sets which have *K* as a subset.

$$I_{\Xi}(K) = P(\Xi \in F_{K\subset *}) = P\{K \subset \Xi\} \text{ where } F_{K\subset *} = \{F \in \Im : K \subset F\} \tag{2-16}$$

The inclusion functional can be used to describe a random set; however, it does not generally, uniquely determine the distribution of a random set due to some pathological cases. It's alternative interpretation is its relation to the capacity functional of $\Xi^{\mathbf{C}}$ [1].

$$I_{\Xi}(K) = P(\Xi^{\mathbf{C}} \cap K = \phi) = 1 - T_{\Xi^{\mathbf{C}}}(K). \tag{2-17}$$

Definition 2.21 Containment functional: The containment functional which calculates the measure of the sets which are contained in *K*.

$$C_{\Xi}(K) = P(\Xi \in F_{*\subset K}) = P\{\Xi \subset K\} \tag{2-18}$$

where $F_{*\subset K} = \{F \in \Im : F \subset K\}$.

It can be shown that the containment functional is completely intersection monotone making it the dual of the capacity functional [1]. It can be shown that the following relationship exists between the capacity and containment functionals:

$$C_{\Xi}(K) = P\{X \subset K\} = 1 - T_{\Xi}(K^{\mathbf{C}}) \tag{2-19}$$

This relation also gives an intuitive explanation as to why the containment functional also determines the distribution of a random set, if defined on the open sets. This dual relationship shared between the capacity and containment functionals is similar to the relationship shared

between belief and plausibility functions. Belief functions are used extensively in evidential reasoning and are discussed in the Theory of Evidence section [8].

For the purposes of the random set, the containment functionals superadditivity property can be viewed as a pessimistic estimate of a random set value. The containment functional uses a containment requirement for the probabilistic frame of reference, meaning it uses sets that are contained in *K* to calculate probability. In other words, this value is the probability that only the elements of *K* will be generated, whereas, the capacity functional requires only the existence of one similar element. In fact, it can be shown that the containment functional is a lower probability

$$C_{\Xi}(K) = \inf\{P(K) : P \in \mathbf{P}_{\Xi}\} \text{ [1].} \tag{2-20}$$

This implies that $C_{\Xi_i}(K)$ is dominated by $P(K)$, $\forall P \in \mathbf{P}_{\Xi}$, $\forall K \in \mathcal{K}$. All probability measures on a random set are wedged in between these bounds, that is

$$C_{\Xi}(K) \le P(K) \le I_{\Xi}(K) + C_{\Xi}(K) + H_{\Xi}(K) = T_{\Xi}(K) \ \forall P \in \mathbf{P}_{\Xi}, \forall K \in \mathcal{K}. \tag{2-21}$$

This is intuitive since the capacity functional is the probability that the random set will hit a given set, whereas the containment functional is the probability that the random set is fully contained within the given set.

Definition 2.22 Hit and miss functional: The hit and miss functional calculates the measure of sets that intersect the set *K*, but have no inclusion or containment relationship.

$$H_{\Xi}(K) = P(\Xi \in F_{*\not\subset K, K\not\subset *}) = P\{\Xi_i \cap K \neq \phi, \Xi \not\subset K, K \not\subset \Xi\} \tag{2-22}$$

where $F_{*\not\subset K, K\not\subset *} = \{F \in \Im : F \not\subset K, K \not\subset F\}$

The hit and miss functional is not used in the literature. It simply identifies sets that have a non-empty intersection with a set *K*, non-containment relationship with a set *K*, and non-inclusion relationship with a set *K*. Its use alone for the purposes of probability assignment would not be intuitive.

The inclusion and containment functionals identify the sets *above* or *below K* in the lattice of subsets of $\Im_K$, that is these functionals identify the sets that can be linearly ordered with respect to *K* by inclusion and containment. On the other hand, the hit and miss functional considers all sets at the same level as *K* on the lattice, and are not comparable using inclusion and containment.

Theory of Evidence

We briefly discuss the relationship between random sets and the *Theory of Evidence*, as developed by Dempster and Shafer.

Definition 2.23 Belief function: A function $BEL: 2^X \rightarrow [0,1]$ is a belief function on some space *X* if the following constraints are satisfied

1. $BEL(\phi) = 0$
2. $BEL(X) = 1$
3. *BEL* is completely monotone [1], [8].

Definition 2.24 Plausibility functions: The dual of the belief function, the plausibility function has the expected dual form

$$PL(A) = 1 - BEL(A^C). \tag{2-23}$$

Just as the capacity functional is an optimistic estimation of the probability of a set outcome, the plausibility function is an optimistic estimation of the probability of an occurrence of an element in *A*. Belief functions are completely determined by their mass functions.

Definition 2.25 Mass functions: A function $m: 2^X \rightarrow [0,1]$ is a mass function if $m(\phi) = 0$ and $\sum_{A \subseteq X} m(A) = 1$.

Note that the containment functional of a random closed set is a belief function, which can also be described by its corresponding mass function.

$$BEL(A) = \sum_{B \subseteq A} m(B) = P_\Xi\{\Xi \subseteq A\} = C_\Xi(A); \tag{2-24}$$

Whereas, a general belief function is a containment functional only if some continuity conditions are met [1].

Note that $\left(2^X, \sigma(2^X), m\right)$ forms a probability space, where *m* is a probability on sets $A \in 2^X$. Furthermore, the corresponding belief function resembles a cumulative distribution function on 2^X using containment relationship to accumulate measure.

The purpose of distributing mass, *m*, to subsets of outcomes rather than simply the outcomes themselves in evidential reasoning is an attempt to model uncertainty. Rather than merely having the ability to state the probability of each outcome, the mass function can assign probability of an outcome occurring in a set without explicitly expressing the probability of its constituents [8].

Point Process

General random set models are seldom used in the machine learning community. This is interesting since random variables and statistical models are ubiquitous in the same community. One reason for this is that the general random set has no simple or even established parametric form or simple methods for estimation. Specific types of random sets, such as point processes, do have simple parametric forms which allow for optimization and estimation; however as will be discussed, they are rarely used to model sets of occurrences.

Next, we define some popular parametric forms of the point process and discuss their pros and cons. We conclude that most parametric forms of the point process are restricted to behave as standard random variables. They do not take advantage of the information attained from the co-occurrence, or observation, of a set of samples, but rather treat these samples as independent occurrences.

Definition 2.26 Counting measure: Assume $\mathcal{E} \subseteq \mathcal{R}^d$ is a topological space. A measure μ on a family of Borel sets $\mathcal{B}(\mathcal{E})$ is called a counting measure if it takes only non-negative integer values, that is $\mu : \mathcal{B}(\mathcal{E}) \to \{0,1,2,...\}$ [4].

A counting measure is locally finite if the measure is finite on bounded subsets of $\mathcal{E}$. Therefore, a locally finite counting measure has a finite number of points in its support in any compact set [4].

Definition 2.27 Point process: A point process $\Phi : \Omega \to N$ is a random closed set with associated probability space, $(\Omega, \sigma(\Omega), P)$, and a measurable space $(N, \mathcal{B}(N))$ where N is the family of all sets φ of points in $\mathcal{E}$ if φ is locally finite (each bounded subset of $\mathcal{E}$ must contain only a finite number of points of φ)[4].

Less formally, a point process is a random choice of $\varphi \in N$ governed by P. In practice, point processes are considered to be random sets of discrete points or as random measures which count the number of points within bounded regions. Random measures are further discussed in the Random Measure section. Since a point process is a random set, the same principles and theorems that apply to random sets apply to point processes.

Since point processes are locally finite, their capacity functional are expressed as follows:

$$T_{\Phi}(K) = \mathbf{P}(\Phi \cap K \neq \phi) = \mathbf{P}(|\Phi \cap K| \neq 0) = \mathbf{P}(\Phi(K) \neq 0), \quad \text{(2-25)}$$

where $\Phi(K) = |\Phi \cap K|$.

Since we know the intersections will have a finite number of elements, we can model these probabilities as counting probabilities [4].

Definition 2.28 Intensity measure: The intensity measure Λ of Φ is the mean value of $\Phi(K)$, defined as $\Lambda(K) = E[\Phi(K)]$, where Λ is simply a random variable with probability space $(\mathcal{K}, \sigma(\mathcal{K}), \mu)$ and measurable space $(\mathcal{R}^{+}, \sigma(\mathcal{R}^{+}))$. Simply, $\Lambda(K)$ is the mean number of points of a realization of Φ in K [1], [4].

In many applications, point processes are modeled in terms of intensity measures to provide for a simpler functional model. It provides for an intuitive idea of intensity and allows for a simple parametric form. The following are examples of a few popular parameterizations: random point, binomial point process, Poisson point process and the Gibbs point process.

Definition 2.29 Random point: A random point is a point process ξ with singleton outcomes. The capacity functional of this random point can be estimated
$P(\{\xi\} \cap K \neq \phi) = P(|\{\xi\} \cap K| \neq 0) = P(\xi \in K)$ [4]. (2-26)

Assume that is our random point is uniformly distributed in some compact set $K \subset \mathcal{E}$. Let ν be the Lesbegue measure on $\mathcal{E}$ that corresponds to length, area, or volume, depending on the dimension of $\mathcal{E}$. Note this measure represents the uniform distribution on the space $\mathcal{E}$. For each subset *A* of *K* we could then define the point process distribution, corresponding to the random point as follows:

$$P(\xi \in A) = \frac{\nu(A)}{\nu(K)} \tag{2-27}$$

This is essentially a standard random variable which should be clear from Equation 2-27. This formulation is simply a ratio of the measure of *A* and the total measure, the measure of *K*. This seems reasonable for the probability of a uniformly distributed random point to fall in volume *A* to assume this value.

Definition 2.30 Binomial point process: A binomial point process with *n* points is *n* independent uniformly distributed random points $\xi_1, \xi_2, ..., \xi_n$ which are distributed over the same compact set $K \subset \mathcal{E}$. This binomial point process, written $\Phi_{W^{(n)}}$ is governed by the following joint distribution

$$P(\xi_1 \in A_1, \xi_2 \in A_2, ..., \xi_n \in A_n) = \prod_{i=1}^{n} P(\xi_i \in A_i) = \frac{\prod_{i=1}^{n} \nu(A_i)}{\nu(K)^n} \tag{2-28}$$

For each subset *A* of *K*. Since ν is a Lebesgue measure, there are three inherent properties of the binomial point process.

1. $\Phi_{W^{(n)}}(\phi) = 0$
2. $\Phi_{W^{(n)}}(K) = n$
3. $\Phi_{W^{(n)}}(A_1 \cup A_2) = \Phi_{W^{(n)}}(A_1) + \Phi_{W^{(n)}}(A_2), A_1 \cap A_2 = \phi$ [4].

The above formulation of random points is indicative of the i.i.d. assumption. The above formulation treats each element of a random set, as being independent of each other. This assumption retards the random sets ability to maintain co-occurrence information about the

samples, and furthermore, behaves similarly to the standard random variable with the i.i.d. assumption.

The aptly named binomial point process has an expected value, $E\left[\Phi_{W^{(n)}}(A)\right]$, modeled by a binomial distribution with parameters *n* and $p = P(\xi \in A)$ [4]. The mean of a binomial distribution is simply the product of its parameters *n* and *p*, yielding

$$E\left[\Phi_{W^{(n)}}(A)\right] = np = \frac{n\nu(A)}{\nu(K)}. \tag{2-29}$$

This means that the intensity—mean number of points per unit volume—is given by

$$\lambda = \frac{n\nu(A)}{\nu(K)}\frac{1}{\nu(A)} = \frac{E\left[\Phi_{W^{(n)}}(K)\right]}{\nu(K)}. \tag{2-30}$$

Although each of the points is distributed uniformly about the sample space in a binomial point process, the number of points contained in subsets of *K*, are not independent, since this distribution is defined for a fixed number of points *n*. If we were to construct $\Phi_{W^{(n)}}$ in terms of the number of points per subset as in [4] , the distribution would be more descriptive.

$$P(\Phi_{W^{(n)}}(A_1) = n_1, ..., \Phi_{W^{(n)}}(A_k) = n_k) \tag{2-31}$$
where $n_1 + n_2 + ... + n_k = n$ and $k = 1,2,...$

Example 2.1 Dependence on number of samples: It is clear that the number of points contained in subsets of *K* are dependent due to the fact that $n_1 + n_2 + ... + n_k = n$. If we know that $\Phi_{W^{(n)}}(A_1) = n_1$, then we also know that $\Phi_{W^{(n)}}(K \setminus A_1) = n - n_1$ [4].

We reiterate that the binomial point process treats its outcomes as the product standard random variables with the i.i.d. assumption and it is highly dependent on the number of points in a given area *A*.

Definition 2.31 Poisson point process: Let Λ be a locally finite measure on a topological space $(\mathcal{E}, \mathcal{B}(\mathcal{E}))$. The Poisson point process Π_Λ with intensity measure Λ is a random subset of $\mathcal{E}$ that satisfies the following constraints
1. For each bounded subset *K* of $\mathcal{E}$, the random variable $|\Pi_\Lambda \cap K|$ has a Poisson distribution

with mean $\Lambda(K)$.

2. Random variables $|\Pi_\Lambda \cap K|$ are independent for disjoint *K* [4].

The corresponding capacity functional takes the form

$$T_{\Pi_\Lambda}(K) = P\{\Pi_\Lambda \cap K \neq \phi\} = 1 - \exp(-\Lambda(K)) \text{ [1], [4].} \tag{2-32}$$

The first constraint suggests that $\Lambda(K)$ is parameterized by λ, the parameter of the Poisson distribution. This parameterization is usually of the form $\Lambda(K) = \lambda\nu(K)$, where ν is a measure, usually Lebesgue, of the set value *K* for all $K \in \mathcal{K}$. The second constraint imposes *independent scattering*, the number of points in disjoint Borel sets are independent. Note that this second constraint implies that there is *no interaction between points in a pattern*—elements in a set [4]. This parameterization would therefore be limiting for context estimation.

The last point process model that is discussed is the Gibbs point process which has roots in statistical physics. They are motivated by Gibbs distributions which describe equilibriums states of closed physical systems. In Gibbs theory, likelihoods of configurations are modeled assuming that the higher the probability of a system of objects, the lower the potential energy of the system [4]. This ideology is modeled in their definition.

Definition 2.32 Gibbs point process: A point process Ψ is a Gibbs Point Process with exactly *n* points if its capacity function is governed by the probability density function defined in Equation 2-33.

$$f(K) = \frac{\exp(-U(K))}{Z} \tag{2-33}$$

Hence the distribution is calculated in the standard fashion.

$$P(\Phi \in K) = \int\cdots\int_K f(x_1,...,x_n)dx_1...dx_n \tag{2-34}$$

In Equation 2-33, the function $U : \mathcal{R}^{nd} \rightarrow \mathcal{R}$ is the energy function and *Z* is the partition function. Note in Equation 2-34, order of integration is irrelevant since $K = \{x_1,...,x_n\}$ [4].

In practice, the energy function is chosen to be a sum of interaction potentials

$$U(K) = \sum_{A \subseteq K} V(A). \tag{2-35}$$

Frequently, V is assumed to have small values for large subsets of K. This assumption leads to the use of a *pair potential* function

$$U(K) = \sum_{i=1}^{n}\sum_{j=1}^{n}\theta\left(\left\|x_i - x_j\right\|\right). \tag{2-36}$$

The Gibbs point process can also be formulated for varying numbers of points n. This is called the *grand canonical ensemble* and assumes n is random [4]. Let $\mathcal{K}_n$ be the family of sets with n points. Then we can define $\mathcal{K} = \bigcup_{n=0}^{\infty}\mathcal{K}_n$ [4].

We can now define a density on $\mathcal{K}$.

$$f(K) = ca^n \exp(-U(K)). \tag{2-37}$$

where c and a are the appropriate normalization factors [4].

Random Measures

Random measures associated with random sets are generalizations of counting measures. As a random counting measure is a function on a point process, a random measure, associated with random sets, is a function on a random set.

Definition 2.33 Random measure: Assume $\mu : \mathcal{F} \to [0,\infty)$ is a fixed measure and Ξ is a random closed set with respect to the Fell Topology. Then $M_{\Xi,\mu}(F) = \mu(F \cap \Xi)$ is a random measure which maps from some probability space $(\Omega, \sigma(\Omega), P)$ to a measurable space $(\mathcal{M}, \mathcal{B}(\mathcal{M}))$ where $\mathcal{M}$ is the family of all locally finite measures on $\mathcal{F}$ and $\mathcal{B}(\mathcal{M})$ is generated by $\{M \in \mathcal{M} : M(F) > t\}$ for every $F \in \mathcal{F}$ and $t > 0$ [1].

For each instance X of Ξ, we have a corresponding instance $M_{X,\mu}$ of random measure $M_{\Xi,\mu}$, specifically a measure taking on a non-negative value for each set F. Note that throughout the literature, the measure μ is assumed to be additive and thus it has all corresponding characteristics. If we restricted $\mu : \mathcal{F} \to [0,1]$, it can define a probability measure on $\mathcal{R}^d$,

namely $P_{\Xi,\mu}(F) = M_{\Xi,\mu}(F)$. Therefore each instance of a random set Ξ has a corresponding measure $P_{X,\mu}$ [1].

$$P_{X,\mu} = \frac{\mu(F \cap X)}{\mu(X)}, \forall F \in \Im \tag{2-38}$$

To avoid cumbersome notation, we may omit μ and refer to $P_{X,\mu}$ as P_X when there is no ambiguity. This construction can be generalized by a taking a measurable random function $\varsigma(x), x \in \mathcal{E}$. We can then define a random measure as in Equation 2-39.

$$M_{\Xi,\mu}(F) = \int_{\Xi \cap F} \varsigma(x) d\mu(x) \tag{2-39}$$

Then we can construct a measure P_X as in Equation 2-40 [1].

$$P_X = \frac{\int_{X \cap F} \varsigma_X(x) d\mu(x)}{\int_X \varsigma_X(x) d\mu(x)}, \forall F \in \Im \tag{2-40}$$

We have therefore defined a mapping from X to $P_{X,\mu}$. Note in this construction we assume a dependence of ς on Ξ, denoted by ς_X.

Note, we have also defined a family of measures P_Ξ associated with random set Ξ. The random measure could be viewed as a distribution on distributions, or a measure on measures, which is related to variational approaches for approximate inference.

Variational Methods

The use of variational methods for approximate inference has become a popular classification method in the machine learning community. We give a brief description in order to identify its relationship to random sets, or more specifically, random measures. The goal of variational approaches is to determine the posterior $P(Z \mid X)$, of latent variables Z given observed data X where Z are typically class labels and parameters of distributions for the

elements of *X*. This approach is typically preferred over standard methods when the latent variable space is large, the expectations with respect to the posterior are intractable, or the integrations required are intractable or have no closed form representation [97].

Variational inference approximation balances the pros and cons of typical estimation approaches such as EM and other more computationally intensive methods such as stochastic techniques [97]. EM approaches suffer from the aforementioned problems; whereas stochastic methods such as Markov Chain Monte Carlo (MCMC) methods can generate exact results, but not in finite time [97].

In standard approaches such as EM, parameters are estimated by inspecting a small portion of the parameter space, which may make it more likely to settle in local optima rather than the global. MCMC methods attempt to construct the true distribution over all the possible values of the parameters using sampling methods. This approach allows for a globally optimal choice of parameter values or allows for the integration over all possible values. However, these approaches are only guaranteed as the sampling tends to infinity, but they may be useful when the sample space allows for a tractable solution [97].

In variational methods for approximate inference, function learning is the objective and typically hyperparameters, prior distributions on a function's parameters, are used to model a family of function values. It can be shown that the optimization of the log likelihood of the set of observations *X* can be separated into two terms:

$$\ln p(X) = L(q) + KL(q \parallel p) \tag{2-41}$$

where $L(q) = \int q(Z) \ln\left(\frac{p(X,Z)}{q(Z)}\right) dZ$ and $KL(q \parallel p) = -\int q(Z) \ln\left(\frac{p(Z \mid X)}{q(Z)}\right) dZ$.

It can also be shown that we can maximize the lower bound *L(q)* by minimizing the KL divergence between $q(Z)$ and $P(Z \mid X)$. Therefore, this is approach is a variational method, as

$p(Z|X)$ is estimated by optimizing the log likelihood with respect to the function q. Given the use of hyperparameters the optimization with respect to q is called a *free form* estimate, that is, q is only restricted by the parameterization of the hyperparameters. Therefore this expression can be seen as the optimization of a functional with respect to a function,

$$H[q] = \ln p(X). \tag{2-42}$$

The parameter distributions are typically formulated for simple integration, such that the parameters can be integrated out for the purposes of inference, usually classification. That is, the parameters are never estimated explicitly.

In summary, variational learning estimates a function through the use of observed data and parameter distributions governed by hyperparameters. These parameter distributions, which are distributions on distributions, are similar to the idea of random measures. However, as discussed in the Technical Approach, the purpose of the random measure within the random set framework is different from the use of hyperparameters in variational inference.

Before we discuss random set applications, it is necessary to review some measures, metrics and divergences defined on sets or measures.

Set Similarity Measures

In data sample analysis, it is necessary to have some sort of similarity measure for the purposes of comparing and contrasting the samples. If we are performing contextual analysis it seems appropriate to have a similarity measure to compare and contrast sets. The following is a brief review of standard and modern set similarity measures.

One way to analyze the similarity of measures would be to use a distribution similarity measure or divergence. Popular examples are the *Kullback-Leibler* (KL) divergence, which was

informally introduced in the previous section, and Chernoff divergence. The well-known KL divergence between distributions P_0 and P_1 is computed as follows:

$$KL(P_1 \,||\, P_0) = \int p_1(x) \log\left(\frac{p_1(x)}{p_0(x)}\right) dx \tag{2-43}$$

The Chernoff divergence is computed as follows:

$$C(P_0, P_1) = \max_{0 \le t \le 1} \left[-\log \mu(t) \right] \tag{2-44}$$
$$\text{where } \mu(t) = \int [p_0(x)]^{1-t} [p_1(x)]^t dx .$$

Upon inspection, both of these divergences seem to quantify the idea of similarity of measures based on the underlying distribution of mass.

Another common approach is the use of compressed distribution similarity measures. Common histogram measures are the L_1 and weight L_2 measures.

$$d_{L1}(H, K) = \sum_i |h_i - k_i| \tag{2-45}$$

$$d^2_{L2}(H, K) = (\mathbf{h} - \mathbf{k})^t \mathbf{A} (\mathbf{h} - \mathbf{k}) \tag{2-46}$$

In Equations 2-45 and 2-46, **A** is a weight matrix; *H* and *K* represent histograms, weighted clusters, or feature subsets of two discrete sets. Although popular, these similarity measures give rise to problems in robustness. For example, when computing the differences in histogram bins, Equations 2-45 and 2-46 do not account for neighboring bins.

A common similarity measure used in topological spaces is the Hausdorff metric. This metric computes the difference between two sets by finding the maximum difference of the minimum point-wise differences.

$$d^H(X, Y) = \max\left(\sup_{x_i \in X} \inf_{y_i \in Y} \|x_i - y_i\|, \sup_{y_i \in Y} \inf_{x_i \in X} \|x_i - y_i\| \right). \tag{2-47}$$

Although this similarity measure is indeed a metric, it seems to lack robustness. For example, two point sets having all constituents the same, less one outlier, would still be assigned a high difference value.

Another recently researched approach is the *earth mover distance* (EMD) [70], [71]. The idea behind the EMD is to calculate the minimum work needed to transform a discrete set *X* into a discrete set *Y* given some constraints. This minimization is done using linear programming. In fact, this distance calculation is a reformulation of the well known transportation problem. In this framework, one of the sets is considered a supplier and one a consumer where each supplier has a supply quantity x_i and each consumer has a demand quantity y_i. Given a shipping cost c_{ij} for each supplier / consumer pair, c_{ij}, the goal is to find the optimal flow of goods, f^*_{ij}, such that the cost is minimal. Using the optimal flow, EMD is calculated as follows:

$$EMD(X,Y) = \frac{\sum_{i\in I}\sum_{j\in J} c_{i,j} f^*_{i,j}}{\sum_{i\in I}\sum_{j\in J} f^*_{i,j}} = \frac{\sum_{i\in I}\sum_{j\in J} c_{i,j} f^*_{i,j}}{\sum_{j\in J} y_j} \tag{2-48}$$

where $\mathbf{f}^* = \arg\min_{\mathbf{f}} \sum_{i\in I}\sum_{j\in J} c_{i,j} f_{i,j}$

subject to

$f_{i,j} \geq 0, \quad \forall i \in I, j \in J,$

$\sum_{i\in I} f_{i,j} = y_j$, and

$\sum_{j\in J} f_{i,j} \leq x_i.$

Note the above formulation requires that each consumer be completely satisfied. For the purposes of set similarity measures, the idea of flow is simply the matching of similar points in the set. The difference between these points is then computed using the cost, which if formulated accordingly, can be a difference measure of these points. Also note that if the numbers of points

are different in the sets *X* and *Y*, then we can assign fraction values to the supplies and demands to allow for fractional point matching.

Houissa *et al.* proposed an algorithm that uses EMD as a metric for the comparison of images for image retrieval from a data base [72]. This is novel approach of using a set metric to analyze the similarity of two sets. In fact, the use of the aforementioned set metrics and divergences is fairly common in the machine learning community.

Random Set Applications

Next, we review current uses of random sets and en masse approaches in the machine learning and pattern recognition communities. The most widely used formulation of the random set is by far the point process [74]-[96].

Point Process Applications

Popular applications of point processes in machine learning and pattern analysis arenas include, but are not limited to, the following: event prediction [89],[90],[92], object recognition / tracking [74], [79]-[83], and particle modeling [4],[85],[93],[94]. Although we do not detail particle modeling, we explicitly mention it since many forms of the point process have deep roots in statistical physics, and therefore, many point process models relate to physics-based concepts. In many fields of physics, one studies the interaction between groups of particles.

In machine learning, these groups of particles are treated as sets of samples distributed by a point process. One of the more popular applications of point processes is event prediction. In this application the point process domain is the real line, typically time, and the particles are events. Other applications include sample clustering. In most applications, the point process is used similarly to standard random variables with standard probabilistic techniques.

There are no known, to the authors, applications of point processes that include the comparison of sets of samples, which is odd since they are random sets. We review some past

and current research involving the use of point processes in a manner relevant to context estimation.

Linnett *et al.* have used Poisson point processes to model segments of images for texture-based classification [84]. In this approach, samples from a same class are considered the same context. Each image is discretized and each pixel with similar gray values is bin grouped into similar point processes. A Bayesian posterior is then calculated estimating the class of each segment. Note that in this approach, the point process is used as a standard clustering algorithm, grouping samples from the same class together.

Stoica *et al.* proposed the Candy model which models road segments, in remotely sensed imagery, as a marked Poisson point process for roadway network extraction [74]. Each line segment is considered a point, or center, with marks such as width, length, and orientation. The interaction of the segments is governed by a Gibbs point process whose energy function contains a data term and a line segment interaction term. The segment interaction term penalizes short line segments. Segments are then merged based on an MCMC sampling method which adds points to segments, deletes points from segments, and merges segments. In later work, they incorporated Gibbs point processes within this model [80].

Descombes, *et al.* used a point process to model segments of images within the Candy model framework [81]. They improved their model by adding a prior density on the line segments. The prior is modeled as a point process, referred to as the Potts model, where the energy function is calculated based on the number of points in a clique in a segment, such that smaller segments are penalized.

Other work, such as extensions of the Candy model, continues their research of the point process for image analysis [82]. They improved their object process which is used to model the

target line networks in remotely sensed images by adding an additional term in its governing density to account for interactions with other object processes.

The point process is used by Savery and Cloutier to model clusters of red blood cells and correlate their orientation with other attributes of the blood [85]. In this paper, the point process is used to model different red blood cell configurations in the presence of backscattering noise. An energy function is used to assign a value to each configuration of blood cells; this function is placed inside an exponential function to estimate the likelihood of each configuration. An MCMC method is then employed to estimate the true configuration of the red blood cells.

En Masse Context-Based Methods

We refer to methods that treat a set of samples as a singleton unit as en masse approaches. These approaches use the same ideology as the random set and attempt to perform inference or analysis using the set.

Dougherty *et al.* proposed a set-based kNN algorithm is proposed to contend with data sets that may be distributed differently with respect to time [12]. In this approach, the idea of context is maintained by using each training set as a set prototype. The algorithm is able to contend with contextual factors and even disguising transformations. In this approach, the *k* nearest neighbors, neighboring training sets, of the test set are identified. Here context is identified through a similarity measure, specifically the Hausdorff metric, between the test *set*, and a prototype set. Classification of the individual samples is performed using the labels of the *k* nearest samples from the *k* nearest sets. Although this approach is improved over other context-based methods and solutions to concept drift, it suffers from a lack of robustness due to the use of the Hausdorff metric.

Bolton and Gader applied set-based kNN to remotely sensed data for target classification [15]. Contextual factors were apparent in this data set. The application of set-based kNN

improved classification results by correctly identifying the contexts using sets of samples; however, the resiliency of the Hausdorff metric was questionable.

Dougherty *et al*. motivated a statistical approach, an extension of set-based kNN, to identify population correlated factors for improved classification [12], [13], [14]. Dougherty *et al*. provided a very theoretical approach which was suggestive of Poisson point processes [12].

We extend Dougherty's theoretical approach and provide a general random set framework for context based classification which permits possibilistic, probabilistic and evidential implementations.

CHAPTER 3
TECHNICAL APPROACH

We propose a context-based approach for classification posed within a random set framework. The incorporation of random sets equips a classification algorithm with the ability to contend with hidden context changes. The goal of the proposed algorithm is, given an input sample set, or population, identify the population's context and classify the individual input samples.

We propose two models for context estimation and provide analogous inference and optimization strategies. The first model is similar to the germ and grain model which is commonly used in point process simulation [4]. We develop possibilistic and evidential approaches within this model and detail some optimization strategies. The second model utilizes random measures. We propose an unnormalized likelihood function which provides for a probabilistic estimate of context within this model. Finally, we provide a discussion to identify the similarities and differences of the proposed random measure model and standard statistical methods.

Mathematical Basis of the Random Set Framework

Assume a topological space $\mathcal{E} = \mathcal{R}^d$ with samples $x \in \mathcal{E}$. Let $\{\Xi_1,...,\Xi_I\}$ be random sets with respect to the Fell topology. Each Ξ_i is used to model a distinct context *i*, where we assume $\{\Xi_1,...,\Xi_I\}$ to be exhaustive. Assume a sample set *X*, test or train, containing a finite number of observations $X = \{x_1, x_2,..., x_n\}$ from some random set. Let $Y : \mathcal{E} \to \mathcal{Z}^+$ be a label function that maps each *x* to a given label $y \in \{1,2,..,l\} \subseteq \mathcal{Z}^+$, where $\mathcal{Z}^+$ denotes the positive integers.

Standard techniques estimate $P(y \mid x)$ for classification. If we believe that *x* was measured or observed in the presence of contextual factors, we can assume that our label function depends

on the context. If Y not independent of some context Ξ in which x was observed, the posterior estimate can be formulated as follows:

$$P(y \mid x, X) = \sum_{i=1}^{I} P(y, \Xi_i \mid x, X) . \tag{3-1}$$

Equation 3-1 is interpreted as calculating the probability that sample x has class label y and was generated in context i. In Equation 3-1 the posterior is marginalized over each potential context i. For reasons developed throughout Chapters 1 and 2, context identification is performed by indentifying contextual transformations; therefore, the observed population X is used for context estimation. Using Bayes' rule and making some independence assumptions, we arrive at Equation 3-2.

$$P(y \mid x, X) = \sum_{i=1}^{I} \frac{P(x, X \mid y, \Xi_i) P(y, \Xi_i)}{P(x, X)} \propto \sum_{i=1}^{I} P(x \mid y, \Xi_i) P(X \mid \Xi_i) P(y \mid \Xi_i) P(\Xi_i) \tag{3-2}$$

In Equation 3-2, we assume x is independent of X given its context and label. We also assume X is independent of y given the context. Equation 3-2 provides a random set framework for context based classification.

The factors in Equation 3-2 have intuitive meanings. The factor $P(x \mid y, \Xi_i)$ can be interpreted as the probability or likelihood that x was collected in context i and is of class y. A suitable implementation would be I classifiers, such that when each is presented with a sample x, could identify it as having class label y in its corresponding context i.

The result of classification within a particular context i, $P(x \mid y, \Xi_i)$, is weighted by the term $P(X \mid \Xi_i)$ which can be interpreted as the probability of observing X in context i. The result is an intuitive weighting scheme that weights each classifier's output based on contextual relevance to the test population.

The $P(y,\Xi_i)$ factor is interpreted as a prior likelihood of observing some class and context. Depending on the implementation, this term may be better estimated using, $P(y,\Xi_i) = P(y \mid \Xi_i)P(\Xi_i)$, where $P(y \mid \Xi_i)$ is the probability of class *y* given context *i* and $P(\Xi_i)$ is the prior probability of context *i*. Note that $P(x \mid y,\Xi_i)$ and $P(X \mid \Xi_i)$ are terms of great interest as they embody the context-based approach and will be further discussed and analyzed.

Estimating $P(x \mid y)$ has been researched for years using various models and estimation techniques. The estimation of $P(X \mid \Xi)$ and $P(x \mid y,\Xi)$ has not been researched quite as thoroughly, especially $P(X \mid \Xi)$. It seems proper that the values $P(X \mid \Xi)$ should be estimated using determining functionals of Ξ. The random set model provides for considerable flexibility since these probabilities can be estimated using evidential, probabilistic, or possibilistic techniques.

The proposed generalized, context-based framework may have different interpretations and a potential myriad of implementations. We develop two models for the estimation of $P(X \mid \Xi)$ within the proposed framework. A germ and grain model is specified and accompanied by possibilistic and evidential approaches for the estimation of $P(X \mid \Xi)$. Then a random measure model is specified and a probabilistic approach is developed for the estimation of $P(X \mid \Xi)$.

Possibilistic Approach

In this possibilistic approach, $P(X \mid \Xi)$ is estimated using the capacity functional.

$$P(X \mid \Xi) = P_{\Xi}(X) = T_{\Xi}(X) \tag{3-3}$$

For the initial development of this model we will let *Y* be a random set. Classification of the samples from *X* can be defined as partitioning the set such that subsets of *X* are assigned some

class label *y*. This first model can be considered a preliminary or intermediate model. The classifier in each context is modeled using the constructs which are modeling the context, that is, Y is a random subset of each Ξ. This possibilistic implementation provides for a simple and efficient parametric model which allows for direct analysis of the driving terms in Equation 3-2 and concurrent optimization of the classifier and contextual parameters. Optimization techniques for classifiers that do not share parameters with the germ and grain model are also provided.

Development

Note that in this initial model we use $P(\{x\} \mid Y, \Xi)$ instead of $P(x \mid Y, \Xi)$. This slight modification is due to the fact that the classifier in this initial implementation is modeled by random set constructs. Therefore the samples must be formally defined as singleton sets. However, this is not always the case and the notation $P(x \mid Y, \Xi)$ should be used, when a standard statistical classifier is used.

For the purposes of analysis, we focus on the terms $P(X \mid \Xi)$ and $P(\{x\} \mid Y, \Xi)$. These terms drive the context-based classifier so their isolation will aid in analysis. We assume the prior probabilities of all contexts $P(\Xi_i)$ are equal and that the probabilities of the class given the context $P(Y \mid \Xi_i)$ are equal. Given this we have

$$P(Y \mid \{x\}, X) \propto \sum_{i=1}^{I} P(\{x\} \mid Y, \Xi_i) P(X \mid \Xi_i) \tag{3-4}$$

We develop a model similar to that of the germ and grain model [4], [5], [16], that is, the random set is modeled as a union of random hyperspheres. This model provides a simple yet versatile parametric model to allow for the estimation of the terms in Equation 3-4. The germs are the random hypersphere centers and the grains refer to the size or volume of the hypersphere, which

is directly related to the radii. If random set Ξ_i follows a germ and grain model, it is defined by Equation 3-5, where ξ_{ij} are the germs and Ξ_{ij} are the grains.

$$\Xi_i = \bigcup_{j=1}^{n_i} (\{\xi_{ij}\} + \Xi_{ij}) \tag{3-5}$$

In Equation 3-5, n_i is the number of grains used to model context *i*. In our model we assume each grain is governed by a random radius r_{ij} that is exponentially distributed.

$$p(r_{ij}) = \lambda_{ij} \exp(-\lambda_{ij} r_{ij}) \tag{3-6}$$

This implies that the probability that $\{x\}$ hits a grain, $P(\{x\} \mid \Xi_{ij})$, can be estimated as follows

$$P(\{x\} \mid \Xi_{ij}) = T_{\Xi_{ij}}(\{x\}) = P(r_{ij} > \left\| x - \xi_{ij} \right\|). \tag{3-7}$$

Substituting the probability density in Equation 3-6 into Equation 3-7 yields

$$P(\{x\} \mid \Xi_{ij}) = 1 - P(r_{ij} \le \left\| x - \xi_{ij} \right\|)) = \exp(-\lambda_{ij} \left\| x - \xi_{ij} \right\|). \tag{3-8}$$

Equation 3-8 is used to model the constituent grains Ξ_{ij} and subsequently used to model Ξ_i and Y. The capacity functional of Ξ_{ij}, $P(\{x\} \mid \Xi_{ij})$, is subsequently used to estimate the capacity functional of Ξ_i.

$$P(X \mid \Xi_i) = P(\Xi_i \in \Im_X) = T_{\Xi_i}(X) \tag{3-9}$$

In this model, the calculation of $P(X \mid \Xi_i)$, follows from the calculation of the capacity functional of the constituent grains.

$$P(X \mid \Xi_i) = 1 - \prod_{j=1}^{n_i} \left(1 - T_{\Xi_{ij}}(X)\right) \tag{3-10}$$

Equation 3-10 states that the probability that *X* hits Ξ_i is the same as the probability that *X* does not miss all $\Xi_{ij}, \forall j = 1,...,n_i$. Given our model, we can calculate $T_{\Xi_{ij}}(X)$ using Equation 3-11.

$$T_{\Xi_{ij}}(X) = \max_{x \in X} T_{\Xi_{ij}}(\{x\}). \tag{3-11}$$

The proof is discussed in the Lemma 3-1.

Lemma 3-1. Let Ξ be a random set taking on set values in $\Im$ and having a probability distribution P_Ξ on $\mathcal{B}(\Im)$ and corresponding capacity functional T_Ξ. If we restrict the elements of $\Im$ to be a random disc or hypersphere then $T_\Xi(X) = \max_{x \in X} T_\Xi(\{x\})$ if X is finite or $T_\Xi(X) = \sup_{x \in X} T_\Xi(\{x\})$ if X is infinite.

Proof. We show if $T_{\Xi_{ij}}(\{x_1\}) \geq T_{\Xi_{ij}}(\{x_2\})$ then $P(\Xi_{ij} \cap \{x_1\} \neq \phi) = P(\Xi_{ij} \cap \{x_1, x_2\} \neq \phi)$, which can we inductively show implies $T_\Xi(X) = \max_{x \in X} T_\Xi(\{x\})$.

Base Case: First assume *without loss of generality* (WLOG) that $T_{\Xi_{ij}}(\{x_1\}) \geq T_{\Xi_{ij}}(\{x_2\})$. If random hypersphere is determined by a random radius, then $P(r \geq d(x_1, c)) \geq P(r \geq d(x_2, c))$, where *d* is some metric, *r* is the radius of the hypersphere and *c* is the hypersphere center. This implies that $d(x_1, c)) \leq d(x_2, c)$ if *r* is governed by a distribution that is monotonic with respect to distance, such as the exponential distribution. This is due to the fact that the probability of intersection is a function of distance only. This implies that each hypersphere that $\{x_2\}$ hits, $\{x_1\}$ must hits. So in this model we can assume

$$T_{\Xi_{ij}}(\{x_1\}) > T_{\Xi_{ij}}(\{x_2\}) \Rightarrow P(\Im_{x_1}) > P(\Im_{x_2}) \Rightarrow \forall K, K \in \Im_{x_2} \Rightarrow K \in \Im_{x_1} \tag{3-12}$$

Equation 3-12 implies that $P(\Xi_{ij} \cap \{x_1\} \neq \phi) = P(\Xi_{ij} \cap \{x_1, x_2\} \neq \phi)$.

Induction Step: Now assume $T_\Xi(K) = \max_{x \in K} T_\Xi(\{x\})$.We show that $P(\Xi_{ij} \cap (K \cup \{x_1\}) \neq \phi) = \max\left(\max_{x \in K} T_\Xi(\{x\}), T_\Xi(\{x_1\})\right)$. We know that there exists some $\hat{x} = \arg\max_{x \in K} T_\Xi(\{x\})$ and therefore $\hat{x} = \arg\min_{x \in K} d(x, c))$, where ties are arbitrarily broken. There are two cases. First assume $d(x_1, c)) \leq d(\hat{x}, c)$, which implies that $T_{\Xi_{ij}}(\{x_1\}) \geq T_{\Xi_{ij}}(K)$. Using the same argument in the Base Case, that is, every hypersphere that hits *K*, must hit $\{x_1\}$. In the other case, if $T_\Xi(\{x_1\}) < T_\Xi(K)$, then by the same logic every hypersphere that hits $\{x_1\}$, must hit *K*. Therefore, $T_\Xi(X) = \max_{x \in K \cup x_1} T_\Xi(\{x\})$ and given the Base Case, is true for all sets of countable size. Thus

$$T_{\Xi_{ij}}(X) = T_{\Xi_{ij}}\left(\bigcup_{x \in X} \{x\}\right) = \max_{x \in X} T_{\Xi_{ij}}(\{x\}). \tag{3-13}$$

Q.E.D.

For classification purposes, assume that some subset of the grains represent some class *Y*, which are identified in some index set C_y.

$$Y = \bigcup_{j:(i,j)\in C_y} (\{\xi_{ij}\} + \Xi_{ij}) \tag{3-14}$$

If we assume that the measure of the random hypersphere overlap in each context, $P(\{x\} \mid \Xi_{ij}, \Xi_{ik}),\ j \neq k$ is negligible, then the term $P(\{x\} \mid Y, \Xi_i)$ can be estimated as follows:

$$P(x \mid Y, \Xi_i) \propto \sum_{j:(i,j)\in C_y} P(\{x\} \mid \Xi_{ij}) . \tag{3-15}$$

The assumption in Equation 3-15 admits simplified update equations during the optimization stage.

Dependent Optimization

In this development, we propose an optimization method that assumes parametric dependence of the classifying and context estimating factors. Optimization of the parameters λ_{ij} is performed using a *minimum classification error* (MCE) objective [86], [87] and [88]. The objective is to maximize the difference between correct and incorrect classification. Equation 3-16 is used as an MCE objective function. Each parameter is updated in an iterative fashion using gradient descent. For optimization purposes, let $X_i \in \mathbf{X} = \{X_1, ..., X_I\}$ be training sets that represent different contexts.

$$D(x, X, \lambda_{ij}) = \begin{cases} -\left(\begin{array}{l} \sum_{j:(i,j)\in C_y} P(\{x\} \mid \Xi_{ij}) P(X \mid \Xi_i) - \\ \sum_{(m,k):(m,k)\notin C_y} P(\{x\} \mid \Xi_{mk}) P(X \mid \Xi_m) \end{array} \right), x \in y \\ \left(\begin{array}{l} \sum_{j:(i,j)\in C_y} P(\{x\} \mid \Xi_{ij}) P(X \mid \Xi_i) - \\ \sum_{(m,k):(m,k)\notin C_y} P(\{x\} \mid \Xi_{mk}) P(X \mid \Xi_m) \end{array} \right), x \notin y \end{cases} \tag{3-16}$$

In Equation 3-16, the second terms sum over context, grain pairs that model a class other than C_y, where C_y is the class modeled by parameter λ_{ij}. This objective can be interpreted as an

optimization of λ_{ij} with respect to observations from the context and class it represents as long as it doesn't hinder the classification of observations from other classes in any context.

For stability and quick convergence, a loss function is used.

$$l(x, X, \lambda_{ij}) = \frac{1}{1+\exp(-D(x, X, \lambda_{ij}))} \tag{3-17}$$

The total loss is then defined by Equation 3-18.

$$L(\lambda_{ij}) = \sum_{X \in \mathbf{X}} \sum_{x \in X} l(x, X, \lambda_{ij}) \tag{3-18}$$

We have the following gradient descent update formula where *t* represents the iteration number and α is the learning rate.

$$\lambda_{ij}^{t+1} = \lambda_{ij}^{t} - \alpha \frac{dL}{d\lambda_{ij}} \tag{3-19}$$

$$\text{where} \frac{dL}{d\lambda_{ij}} = \sum_{X \in \mathbf{X}} \sum_{x \in X} l(x, X, \lambda_{ij})(1 - l(x, X, \lambda_{ij})) \frac{dD}{d\lambda_{ij}}, \tag{3-20}$$

and

$$\begin{aligned}
&\frac{dD}{d\lambda_{ij}} = \left(-\left\|x - \xi_{ij}\right\|\right)\exp(-\lambda_{ij}\left\|x - \xi_{ij}\right\|)P(X \mid \Xi_i) + \\
&\left(\sum_{m \in C} \exp(-\lambda_{im}\left\|x - \xi_{im}\right\|)\right)\left(-\prod_{m \neq j} 1 - \exp(-\lambda_{im}\left\|x^{im} - \xi_{im}\right\|)\right) \times \\
&\left(-\left\|x^{ij} - \xi_{ij}\right\|\right)\left(-\exp(-\lambda_{ij}\left\|x^{ij} - \xi_{ij}\right\|)\right) - \\
&\left(\sum_{m \notin C} \exp(-\lambda_{im}\left\|x - \xi_{im}\right\|)\right)\left(-\prod_{m \neq j} 1 - \exp(-\lambda_{im}\left\|x^{im} - \xi_{im}\right\|)\right) \times \\
&\left(-\left\|x^{ij} - \xi_{ij}\right\|\right)\left(-\exp(-\lambda_{ij}\left\|x^{ij} - \xi_{ij}\right\|)\right)
\end{aligned} \tag{3-21}$$

where $x^{ij} = \arg\max_{x}(P(\{x\} \mid \Xi_{ij}))$.

The germs are not optimized in the experiments. However, similar gradient descent methods could be employed.

The proposed updates indicated by Equations 3-18, 3-19 and 3-20 have the added benefit of concurrently updating classification and contextual parameters since both are implemented as

the same structures. Next, we provide a general optimization strategy using the germ and grain model with a possibilistic estimate. That is, we optimize the contextual parameters based on their ability to correctly estimate context.

Independent Optimization

We estimate the contextual parameters using the following MCE objective.

$$D(\lambda_{ij}) = \sum_{X \in \Xi_i} P(X \mid \Xi_i) - \sum_{X \notin \Xi_i} P(X \mid \Xi_i) \tag{3-22}$$

The objective in Equation 3-22 is to maximize the difference between correct and incorrect context estimation. Using a similar gradient descent strategy, we arrive at Equation 3-23.

$$\begin{aligned} \frac{dD}{d\lambda_{ij}} = & \sum_{X \in \Xi_i} \left(-\prod_{m \neq j} 1 - \exp(-\lambda_{im} \left\| x^{im} - \xi_{im} \right\|) \right) \left(- \left\| x^{ij} - \xi_{ij} \right\| \right) \left(- \exp(-\lambda_{ij} \left\| x^{ij} - \xi_{ij} \right\|) \right) - \\ & \sum_{X \notin \Xi_i} \left(-\prod_{m \neq j} 1 - \exp(-\lambda_{im} \left\| x^{im} - \xi_{im} \right\|) \right) \left(- \left\| x^{ij} - \xi_{ij} \right\| \right) \left(- \exp(-\lambda_{ij} \left\| x^{ij} - \xi_{ij} \right\|) \right) \end{aligned} \tag{3-23}$$

Equation 3-23 provides for efficient optimization of the contextual parameter λ_{ij}, based on maximizing the separation between correct and incorrect contextual identification.

Evidential Model

In the possibilistic approach, we estimate $P(X \mid \Xi_i)$ using the capacity functional. In the evidential approach we use the inclusion functional to estimate the term $P(X \mid \Xi_i)$. There are two major reasons why we have chosen the inclusion functional for evidential modeling rather than the containment functional. First, we have a continuous model with discrete observations. This means the probability of containment would be zero for essentially all possible discrete observations *X*. Second, the inclusion functional is more intuitive for set-valued random elements, whereas, containment, similar to the idea of belief, is intuitive for modeling uncertainty with singleton random elements.

Development

We develop the evidential approach using the germ and grain model and assume the radii are exponentially distributed. Given these assumptions, we calculate the probability of inclusion given one random hypersphere as follows:

$$P\{X \subset \Xi_{ij}\} = P(\{F : X \subset F\}) = \exp(-\lambda_{ij} \left\| \hat{x}^{ij} - \xi_{ij} \right\|), \quad \text{where } \hat{x}^{ij} = \arg\min_{x}(P(\{x\} \mid \Xi_{ij})). \tag{3-24}$$

For calculation of inclusion, note that we use $\hat{x}^{ij}$ rather than x^{ij}. As x^{ij} is the closest $x \in X$ to germ ξ_{ij} and determines a non-empty intersection relationship of X and Ξ_{ij}, $\hat{x}^{ij}$ is the furthest $x \in X$ to germ ξ_{ij} and determines an inclusion relationship of X and Ξ_{ij}.

This probability can be accumulated across the constituent random hyperspheres using the same ideology taken during the calculation of the capacity functional in Equation 3-10. Therefore we calculate the probability of inclusion of random set Ξ_i across the constituent hyperspheres using Equation 3-25.

$$P(X \mid \Xi_i) = P\{X \subset \Xi_i\} = P(\{F : X \subset F\}) = 1 - \prod_{j=1}^{n_i} \left(1 - \exp(-\lambda_{ij} \left\| \hat{x}^{ij} - \xi_{ij} \right\|)\right). \tag{3-25}$$

Equation 2-25 states that the probability that a random set Ξ_i includes a set *X* is equal to the probability that each of the constituent random hyperspheres Ξ_{ij}, *does not* have a *non-inclusion* relationship with *X*.

Optimization

Using the objective defined in Equation 3-22, the parameters can be optimized using gradient descent as defined in Equation 3-19. For the optimization of λ_{ij}, we substitute Equation 3-26 into Equation 3-19.

$$\frac{dD}{d\lambda_{ij}} = \sum_{X \in \Xi_i} \left(-\prod_{\substack{m=1 \\ m \neq j}}^{n_i} \left(1 - \exp(-\lambda_{im} \left\| \hat{x}^{im} - \xi_{im} \right\|)\right) \right) \left(- \exp(-\lambda_{ij} \left\| \hat{x}^{ij} - \xi_{ij} \right\|)\right) \left(- \left\| \hat{x}^{ij} - \xi_{ij} \right\|\right) -$$
$$\sum_{X \notin \Xi_i} \left(-\prod_{\substack{m=1 \\ m \neq j}}^{n_i} \left(1 - \exp(-\lambda_{im} \left\| \hat{x}^{im} - \xi_{im} \right\|)\right) \right) \left(- \exp(-\lambda_{ij} \left\| \hat{x}^{ij} - \xi_{ij} \right\|)\right) \left(- \left\| \hat{x}^{ij} - \xi_{ij} \right\|\right). \quad (3\text{-}26)$$

Note we have performed optimization independent of the classifier which is assumed to be independent of λ_{ij}. Depending on the classifier utilized, similar optimization techniques could be used for its parameters.

Probabilistic Model

In the probabilistic approach, we model context using a class of functions on random sets called random measures. That is, for each observed set we construct a corresponding measure. We perform analysis in this space of measures rather than in closed subsets of $\mathcal{E}$, or $\Im$, as in previous models, in hopes of extracting supplementary information to that found during analysis in $\Im$.

Development

Recall in Equation 2-33, a likelihood function was derived for a Gibbs point process using an energy function U which was used to assign likelihood based on the configuration of points in some set *X*. We have noted that different forms of *U* yield different issues and may imply certain constraints on a point process.

We now define an unnormalized likelihood function using an energy functional which calculates the energy of a particular configuration by analyzing an observed function or measure. The goal is to permit a tractable contextual estimate, as opposed to an energy function as in Equation 2-35. Furthermore we desire the ability to analyze the *shape* of a function across $\mathcal{E}$ rather than inspecting pairs of elements in $\mathcal{E}$ as in Equation 2-36. Also, we define the likelihood

function such that it can be parameterized to recognize different random measures, whereas Gibbs point processes are typically used to calculate probability using the energy of a closed system and not necessarily distinct random measures' characterizations.

Since we are analyzing functions, we use the KL divergence on functions. We note that other measures or divergences on functions may be used as well. We define the energy functional for random measure M_{Ξ} as

$$U_{\Xi}(P_X) \equiv KL(P_X \parallel Q_{\Xi}). \tag{3-27}$$

We refer to Q_{Ξ} as the *representative measure* for random measure M_{Ξ} and it can be thought of as a parametric representation of Ξ. We can now define the unnormalized likelihood functional for random measure M_{Ξ} as

$$p_{M_{\Xi}}(P_X) \equiv \exp\left(-KL(P_X \parallel Q_{\Xi})\right). \tag{3-28}$$

Note that this likelihood compares how measure is distributed between the function P_X and Q_{Ξ}. Hereafter, we denote Q_{Ξ} by Q or Q_i for a particular context *i*. If the distribution of mass in P_X becomes more similar to that in Q, a higher likelihood is assigned to P_X, using the KL divergence to assess similarity. Therefore, an intuitive value for Q would be the measure that minimizes the sum over the KL divergences of observed samples $D = \{P_{X_1}, P_{X_2}, ..., P_{X_n}\}$ from M_{Ξ_i},

$$Q \equiv \underset{R \in \mathcal{M}}{\arg\inf}\left(\sum_{j=1}^{N} KL\left(P_{X_j} \parallel R\right)\right). \tag{3-29}$$

Hereafter, we denote the densities corresponding to measures *Q* and P_X as *q* and ν_X respectively, and assume they exist. The likelihood function defined in Equation 3-28 is used for contextual estimation given the random set framework for context-based classification.

Specifically, we use the likelihood on random measures to calculate the contextual estimation term.

$$P(X \mid \Xi_i) \propto \exp\{-KL(P_X \parallel Q_i)\} \tag{3-30}$$

In Equation 3-30, Q_i is the representative measure for context *i* and P_X is the measure corresponding to observed set *X*. We use the KL divergence to compare distributions using their corresponding densities ν_X and q_i, to determine the likelihood of context *i*. Therefore, we can calculate or approximate Equation 3-30 using Equations 3-31 or 3-32, respectively.

$$P(X \mid \Xi_i) = P(\nu_X \mid q_i) = \exp\left\{-\int \nu_X(x) \log\left(\frac{\nu_X(x)}{q_i(x)}\right) dx\right\} \tag{3-31}$$

$$P(X \mid \Xi_i) = P(\nu_X \mid q_i) \approx \exp\left\{-\sum_{x \in A} \nu_X(x) \log\left(\frac{\nu_X(x)}{q_i(x)}\right) \Delta x\right\} \tag{3-32}$$

In Equation 3-32, $A \subseteq \mathcal{E}$ is used to estimate the KL divergence. The choice of *A* is further detailed in the Discussion section.

The choice between Equations 3-31 and 3-32 depends on the formulation of parameter q_i, specifically, whether an analytical representation of the KL divergence exists or whether it is convenient for parameter estimation given an assumed parametric form of the model.

The density q is the parameter for $P(\nu_X \mid q)$, which itself may be parameterized for convenience, for example, $q \sim N(\mu, \Sigma)$ or $q \sim Exp(\lambda)$. We note that estimation may benefit if density q is modeled using a more complex distribution such as a Gaussian mixture; however, this may lead to difficulty in computation and may complicate parameter learning [98].

In the probabilistic approach, we need to construct the ν_X given some observed set *X*. One possible construction would be to use a simple Lesbegue or uniform measure over the discrete points in *X*.

Example 3.1 Discrete measure: Assume $X = \{x_1, x_2, ..., x_n\}$. Then we could construct our measure M_X using a cardinality based measure μ_c such that

$$P_X(F) = \frac{\mu_c(F \cap X)}{\mu_c(X)} = \frac{|F \cap X|}{|X|}. \tag{3-33}$$

We note that this measure meets the requirements outlined in the definition of a random measure; however, it is discontinuous, not smooth, which may lead to optimization issues. Furthermore, as we will see during the construction, issues may arise if M_X has a limited support. Therefore, it is beneficial to provide a parametric measure to provide a smooth measure with a large support.

If we use the generalized development of the random measure, and therefore the general construction of an instance of a random measure as in Equation 2-40, we can develop a parametric measure that is continuous and has a large support, given some assumptions.

Example 3.2 Continuous parametric measure: Assume $X = \{x_1, x_2, ..., x_n\}$ are a finite number of observations from some infinite set $\mathcal{X} \subseteq \mathcal{R}^d$. If we assume that elements in *X* are similarly distributed to this continuous set in space $\mathcal{R}^d$ we could estimate the measure on this set using parameters calculated from *X*, and define a measure $M_X : \mathcal{F} \to [0,1]$, by

$$P_X(F) = \frac{\sum_{X \cap F} \varsigma_X(x)\Delta x}{\sum_{X} \varsigma_X(x)\Delta x} = \frac{\sum_{X \cap F} N(x \mid \mu_X, \Sigma_X)\Delta x}{\sum_{X} N(x \mid \mu_X, \Sigma_X)\Delta x} \approx \frac{\int_{x \in F \cap \mathcal{R}^d} N(x \mid \mu_X, \Sigma_X)dx}{\int_{x \in \mathcal{R}^d} N(x \mid \mu_X, \Sigma_X)dx}. \tag{3-34}$$

We estimate the center of mass μ_X and covariance function Σ_X of the set $\mathcal{X}$ using the set of observed finite samples in *X* and use these estimates for the parameters of the Gaussian density. We have therefore constructed an example of a measure given an observed sample *X*, which is continuous, has a large support, and has a parametric form.

Other parametric forms of ν_X could be developed through many existing methods. If we assume a complicated parametric form for ν_X, some methods that might be used to estimate ν_X, such as the standard EM algorithm, may be subject to initialization conditions and therefore will not strictly satisfy Equation 2-40.

Optimization

Next, we develop optimization strategies and example model implementations that would use Equation 3-31 or 3-32. The developed probabilistic model allows for closed form solutions for optimization given certain model assumptions and appropriate objective functions. Roughly

speaking, the optimization of parameter q, using parametric representations of ν_X, proceeds in two main steps. During the first step, parameters of the densities ν_X are estimated for each $X \in \mathbf{X}$, using standard methods such as EM or ML estimates. The result is a set of densities, and therefore measures, $\{\nu_{X_1}, \nu_{X_2}, ..., \nu_{X_n}\}$. In the second step, representative measure q_i is estimated for each random set Ξ_i by maximizing a likelihood function that is a product of factors involving context dependent classification factor $p(x \mid y, \Xi_i)$, context estimation factor $P(X \mid \Xi_i) = P(\nu_X \mid q_i)$, and prior $P(\Xi_i)$, with respect to function q_i. We focus on the maximization of $P(\nu_X \mid q_i)$ since the classification factors of each context can be estimated using standard techniques. Note that factor $P(\nu_X \mid q_i)$ treats ν_X as the samples rather than *x* as in standard methods.

Specifically in the first optimization example, we assume a form of ν_X and q such that the integral in Equation 3-31 can be calculated analytically. We take an EM approach for optimization; specifically we take an expectation over the contextual parameters given each ν_X constructed from observation set *X*. We assume $q \sim N(\mu, \Sigma)$ and $\nu_X \sim N(\mu, \Sigma)$. Initially each ν_X is constructed from the observed samples from the corresponding $x \in X$. Once each ν_X is constructed, the individual elements of the sets $x \in X$ are no longer referenced in the optimization process.

We begin by defining our objective and corresponding log likelihood function given our initial independence assumptions of the random set framework arriving at

$$L(\Theta) = \log\left(\prod_{X \in \mathbf{X}} P(X \mid \Xi_i) P(\Xi_i) \prod_{x \in X} p(x \mid y, \Xi_i) \right). \tag{3-35}$$

Next, we take an expectation over the contextual parameters given our observed populations,

$$\underset{\Xi|X}{E}\left[L(\Theta)\right] = \sum_{X\in\mathbf{X}}\sum_{i=1}^{I}\left[\log\left(P(X\mid\Xi_i)\right)+\log\left(P(\Xi_i)\right)+\sum_{x\in X}p(x\mid y,\Xi_i)\right]P(\Xi_i\mid X)\,. \tag{3-36}$$

We disregard the classification term for now, as this type of optimization is ubiquitous throughout the literature, and therefore we focus on the contextual terms.

$$R(\Theta) = \sum_{X\in\mathbf{X}}\sum_{i=1}^{I}\left[\log\left(P(X\mid\Xi_i)\right)+\log\left(P(\Xi_i)\right)\right]P(\Xi_i\mid X) \tag{3-37}$$

Using Equation 3-31, we get

$$R(\Theta) = \sum_{X\in\mathbf{X}}\sum_{i=1}^{I}\left[\log\left(\exp\{-KL(\nu_X\parallel q_i)\}\right)+\log\left(P(\Xi_i)\right)\right]P(\Xi_i\mid X)\,. \tag{3-38}$$

After some algebra we arrive at

$$R(\Theta) = \sum_{X\in\mathbf{X}}\sum_{i=1}^{I}\left[-\int\nu_X(x)\log\left(\frac{\nu_X(x)}{q_i(x)}\right)dx+\log\left(P(\Xi_i)\right)\right]P(\Xi_i\mid X). \tag{3-39}$$

Analytically integrating and ignoring a constant [98], we arrive at

$$R(\Theta) = \sum_{X\in\mathbf{X}}\sum_{i=1}^{I}\left[-.5\left(\log\left(\frac{\left|\Sigma_{q_i}\right|}{\left|\Sigma_{\nu_X}\right|}\right)+Tr\left[\Sigma_{q_i}^{-1}\Sigma_{\nu_X}\right]+(\mu_{\nu_X}-\mu_{q_i})^T\Sigma_{q_i}^{-1}(\mu_{\nu_X}-\mu_{q_i})\right)+\right.$$
$$\left.\log\left(P(\Xi_i)\right)\right]P(\Xi_i\mid X) \tag{3-40}$$

We then perform the maximization step by differentiating Equation 3-40 with respect to the parameters. At this point we note that many closed form representations can be found for the KL divergence of distributions other than the Gaussian, such as the exponential distribution. Setting the result of the differentiation of Equation 3-40 to zero and solving for parameters μ_{q_i}, Σ_{q_i}, and $P(\Xi_i)$, results in update Equations 3-41, 3-42, and 3-43, respectively.

$$\mu_{q_i} = \frac{\sum_{X \in \mathbf{X}} \mu_{\nu_X} P(\Xi_i \mid X)}{\sum_{X \in \mathbf{X}} P(\Xi_i \mid X)} \tag{3-41}$$

$$\Sigma_{q_i} = \frac{\sum_{X \in \mathbf{X}} \left[\Sigma_{\nu_X} + (\mu_{\nu_X} - \mu_{q_i})(\mu_{\nu_X} - \mu_{q_i})^T\right] P(\Xi_i \mid X)}{\sum_{X \in \mathbf{X}} P(\Xi_i \mid X)} \tag{3-42}$$

$$P(\Xi_i) = \frac{\sum_{X \in \mathbf{X}} P(X \mid \Xi_i)}{\sum_{i=1}^{I} \sum_{X \in \mathbf{X}} P(X \mid \Xi_i)} \tag{3-43}$$

Finally, we use Bayes' rule to solve for $P(\Xi_i \mid X)$.

$$P(\Xi_i \mid X) \propto P(X \mid \Xi_i) P(\Xi_i) \tag{3-44}$$

Recall, $P(X \mid \Xi_i)$ is given by Equation 3-31.

However as previously mentioned, if a more complex distribution is assumed for the model or the sample ν_X, the KL divergence may not have a closed form representation. We now develop an optimization strategy for this case.

Assume the representative measure is a Gaussian mixture, $q_i \sim \sum_{j=1}^{n_i} N(\mu_{ij}, \Sigma_{ij})$, which does not permit a closed form solution. We note there are numerical / statistical methods that can be used to help estimate the KL divergence [98]; however, the optimization of the parameters in q_i would become an issue if those techniques were used.

For development of this optimization technique, we skip to Equation 3-37 and substitute in Equation 3-32 arriving at

$$R(\Theta) = \sum_{X \in \mathbf{X}} \sum_{i=1}^{I} \left[-\sum_{x \in A} \left[\nu_X(x) \log\left(\frac{\nu_X(x)}{q_i(x)} \right) \Delta x \right] + \log\left(P(\Xi_i)\right) \right] P(\Xi_i \mid X). \tag{3-45}$$

Upon inspection, we see that optimization with respect q_i is analogous to minimizing the KL divergences between each ν_X and q_i. If we assume q_i is a Gaussian mixture, with some algebra, we arrive at

$$R(\Theta) = \sum_{X \in \mathbf{X}} \sum_{i=1}^{I} \left[-\sum_{x \in A} \left[\nu_X(x) \log\left(\nu_X(x)\right) \Delta x - \nu_X(x) \log\left(\sum_{j=1}^{J} \pi_{ij} N(x \mid \mu_{ij}, \Sigma_{ij}) \right) \Delta x \right] + \log\left(P(\Xi_i)\right) \right] P(\Xi_i \mid X). \tag{3-46}$$

After performing the maximization step for parameter μ_{ij} we can get a closed form solution assuming Equation 3-48 is independent of μ_{ij}.

$$\mu_{ij} = \frac{\sum_{X \in \mathbf{X}} \left[\sum_{x \in A} \left[x[\Delta x \nu_X(x)] \gamma_{xij} \right] \right] P(\Xi_i \mid X)}{\sum_{X \in \mathbf{X}} \left[\sum_{x \in A} \left[\Delta x \nu_X(x) \gamma_{xij} \right] \right] P(\Xi_i \mid X)} \tag{3-47}$$

$$\text{where } \gamma_{xij} = \frac{\pi_{ij} N(x \mid \mu_{ij}, \Sigma_{ij})}{\left(\sum_{j=1}^{J} \pi_{ij} N(x \mid \mu_{ij}, \Sigma_{ij}) \right)} \propto p(\mu_{ij} \mid x) \tag{3-48}$$

While updating the parameters, we assume γ_{xij} is independent of the other parameters, which is a common assumption in machine learning [97]. In fact, this result is a similar to the result attained using a standard EM approach, taking the expectation over each component given the individual samples using $p(\mu_{ij} \mid x)$ [97]. The other parameters are solved similarly,

$$\Sigma_{ij} = \frac{\sum_{X \in \mathbf{X}} \left[\sum_{x \in A} (x - \mu_{ij})(x - \mu_{ij})^T \left[[\Delta x \nu_X(x)] \gamma_{xij} \right] \right] P(\Xi_i \mid X)}{\sum_{X \in \mathbf{X}} \left[\sum_{x \in A} \left[\Delta x \nu_X(x) \gamma_{xij} \right] \right] P(\Xi_i \mid X)} \tag{3-49}$$

$$\text{and } \pi_{ij} = \frac{\sum_{X \in \mathbf{X}} \left[\sum_{x \in A} \left[\Delta x \nu_X(x) \gamma_{xij} \right] \right] P(\Xi_i \mid X)}{\sum_{X \in \mathbf{X}} \left[\sum_{x \in A} \left[\Delta x \nu_X(x) \right] \right] P(\Xi_i \mid X)}. \tag{3-50}$$

Optimization is again performed in sequence with parameter γ_{xij} being calculated last in each epoch.

To properly calculate the factor Δx in the update Equations 3-47, 3-49 and 3-50, we use the standard approximation of the Riemann integral. If *x* is multidimensional, $x \in \mathbb{R}^d$, construction of Δx involves creating incremental volumes ΔV by constructing a hypergrid or hyper-rectangles. Hereafter, we refer to ΔV as Δx. One intuitive method of constructing the set *A* would be to construct samples by taking all N^d combinations of the *N* samples in *X* in each dimension *d*. However, if samples *x* are multidimensional, then construction of Δx may be intractable. If a smaller *A* was constructed, the Riemann approximation may decrease in accuracy.

We propose an efficient estimation of the KL divergence that assumes Δx is constant and that the samples that comprise *A* are uniformly sampled from some hyperrectangle created from observations of the distributions ν_X and q_i. This approximation, which is similar to Markov Chain Monte Carlo (MCMC) integration, is intuitive since if the samples are, in fact, uniformly distributed, Δx should be constant. In the experiments, we analyze the error using synthetic and real data sets.

Discussion

There are many interesting results of this derivation. For clarification, we first provide a few examples in order to flesh out some of these details. Next we discuss certain similarities and distinctions between the proposed method, standard methods and variational methods. In particular we compare optimization and inference results of the proposed method to standard statistical methods. Lastly, we compare the proposed method to typical variational methods for approximate inference.

We noted earlier that in the construction of the proposed likelihood function we hoped to gain some versatility over standard probabilistic approach that assume i.i.d. However, some approaches that could be employed for the construction of ν_X, may implicitly assume that the singleton elements of *X* are i.i.d. However, we note these effects do not necessarily trickle up to inference at the measure level. After we introduce the optimization methods, which helps identify some characteristics of the proposed approach, we illustrate some of the similarities and differences between using standard statistical approaches which assume i.i.d. and the proposed method.

Example 3.3 Construction of ν_X : Equation 3-32 can be rewritten

$$P(\nu_X \mid q_i) = \prod_{x \in A} \exp\left\{ -\nu_X(x) \log\left(\frac{\nu_X(x)}{q_i(x)} \right) \Delta x \right\}. \tag{3-51}$$

Note that that ν_X is a function of our observation set, *X* and therefore each term in Equation 3-33 is dependent on the set *X*. Note, the use of samples $x \in A$ is simply to estimate the KL divergence, that is, the only reason to use the underlying space is to sample the values of ν_X and q_i. In fact, the samples in *A* do not even need to be elements of the observation set *X*.

Since the likelihood function can be factorized as in Equation 3-33, we could interpret the resulting product as stating that each value $\nu_X(x)$ is distributed by standard random variable $M_{\Xi_i}(x)$ which is determined by random set Ξ_i and is represented by $q_i(x)$ given representative function q_i. Note that ν_X is a function of the set *X* and that each corresponding value $\nu_X(x)$ is drawn from a *distinct* random variable $M_{\Xi_i}(x)$ at each *x* in the domain of M_{Ξ_i}, as illustrated in Figure 3-1.

So in effect, a random measure is a continuum of random variables on some subset of $\mathcal{E}$, one for each element in the domain of M_Ξ, namely $M_\Xi(x)$. As mentioned, each random variable $M_\Xi(x)$ has a corresponding parameter $q(x)$. Note the parametric form of the representative function q, has allowed us to maintain a continuum of random variables in a concise manner, but at the cost of versatile forms of $q(x)$ and therefore $M_\Xi(x)$. That is, $M_\Xi(x_1)$ is a random variable that maps into $\mathcal{R}$ whose distribution is intrinsically governed by

the distribution of the random set Ξ. We shall refer to the value $q(x_1)$ as the representative value for random variable $M_\Xi(x_1)$.

Example 3.4 Random variable $M_\Xi(x)$: If we wanted to minimize the KL divergence of two probability measures, the two functions must coincide. Assume we wanted to minimize the sum of KL divergences between *q* and samples ν_X. At the optima each representative value $q(x)$ is the representative function's value at *x* which minimizes this sum of the KL divergences, given the constraint that the representative function must be probability measure. Assume we collect *N* samples from Ξ and have *N* corresponding measures. Note that given for each instance $\nu_{X_j}(x), j = 1,..., N$ is an instance of random variable $M_\Xi(x)$ for a fixed *x*. If we minimize the expression $\sum_j KL\left(P_{X_j} \| Q\right)$ with respect to each value $q(x)$ at a fixed *x*, using Equation 3-32 and subject to the constraint that *Q* must be a probability measure, that is $\sum_{x\in A} q(x)\Delta x = 1$, we arrive at

$$q(x) = \frac{\sum_j \nu_{X_j}(x)}{\sum_{x\in A}\sum_j \nu_{X_j}(x)\Delta x} = \frac{\sum_j \nu_{X_j}(x)}{\sum_j\sum_{x\in A} \nu_{X_j}(x)\Delta x} = \frac{\sum_j \nu_{X_j}(x)}{\sum_j 1} = \frac{\sum_j \nu_{X_j}(x)}{N}. \quad (3\text{-}52)$$

Note that $q(x)$ is the arithmetic mean of $\nu_{X_j}(x), j = 1,..., N$. This means the *representative value is the mean value for $M_\Xi(x)$ for each x, and therefore minimizes the squared Euclidean distance* between samples $P_X(x)$ from random variable $M_\Xi(x)$, as illustrated in Figure 3-1B.

One result of using a parametric form for the representative function is that the representative value $q(x)$ may no longer be the exact mean of random variable $M_\Xi(x)$ due to the particular constraints, for example if it is assumed Gaussian distributed. However, the assumption of a parametric model is important, otherwise, we would need a random variable for each point in the domain of M_Ξ which does not permit a tractable solution, unless a very simple domain is assumed. As found throughout machine learning techniques, there is a tradeoff between data fidelity and tractability.

Example 3.5 Representative value: Given a set of observed Gaussian measures constructed by selecting the mean and covariance from a uniform interval, assume we wish to construct *Q* using the update Equations 3-47, 3-49 and 3-50. Note this implies we are assuming that *q* is Gaussian. The resulting representative values are not necessarily the arithmetic means of samples from $\nu_\Xi(x)$, as illustrated in Figure 3-1B. Although, the update equations are optimal

assuming Gaussian, there are not necessarily optimal over all possible distributions due to this extra constraint.

The first optimization technique proposed that uses Equation 3-31 to calculate the KL divergence, learns the parameters of *q* using some parameters of our observed distribution ν_X. However, in the second optimization technique proposed, the parameters of *q* are learned using the underlying space, samples in $\mathcal{R}^d$. We note there is some similarity between these update equations and those that are developed in standard EM algorithms such as Equation 3-53.

$$\mu_{ij} = \frac{\sum_{x \in A} x \gamma_{xij}}{\sum_{x \in A} \gamma_{xij}}. \tag{3-53}$$

However, we note that in the proposed update equations, there is a discrete expectation over random measures, not simply an expectation over standard random variables. We also note that when samples are clustered, the set *A* is typically the data. However, in the proposed approach, the samples in set *A* are not directly important, as long as their use permits a good estimate of the KL divergence.

The major difference in the update formulas is the factor $\nu_X(x)\Delta x$. Note that in the KL divergence we integrate with respect to our sample ν_X, which is also a density. In the discrete approximation, the factor $\nu_X(x)\Delta x$ is used instead. One interpretation is that we are taking the expected value of the difference between ν_X and q_i. This interpretation shows, that during optimization, we are trying to minimize the difference between samples ν_X and representative measure q_i.

Another interpretation is the representative function q_i is being coerced into a form similar to the samples ν_X vicariously through its parameters μ and σ using samples *x* in *A* and weights

$\nu_X(x)\Delta x$, as illustrated in Figure 3-2. This coercion is performed through the parameters, for example, the means μ, which reside in the same space as *x*. For this reason, the samples *x* are included in the update equations. However in Equation 3-47, the factor $\nu_X(x)\Delta x$ weights each sample *x* by its corresponding measure in the distribution, ν_X. In fact, μ is optimized such that $q(x)$ is similar to $\nu_X(x)$, not necessarily to maximize $q(x)$ with respect to the samples *x* as is the case with standard statistical optimization.

However, there are similarities in standard statistical optimization and the proposed method. In standard statistical methods, the learned posteriors / likelihoods are optimized while assuming i.i.d. In the proposed method, the representative function is optimized using observed measures which may have been constructed using similar optimization techniques that are used in standard statistical methods. In the developed approach, the observed measures are essentially likelihood functions optimized with respect to each observed set, and therefore most likely assume i.i.d. during their construction. We illustrate situations when the proposed methods result in similar and different optimizations than standard methods.

Example 3.6 Optimization similarities and differences: Assume we have multiple observation sets $D = \{X_1, X_2, ..., X_N\}$ observed in the same context and we wish to optimize likelihood functions for context estimation. We optimize a standard likelihood function, which assumes i.i.d., using the EM algorithm while training on the dataset $X = \bigcup_{i=1}^{N} X_i$. We will also learn the proposed random measure likelihood function by optimizing the representative function given observed measures $\{\nu_{X_1}, \nu_{X_2}, ..., \nu_{X_N}\}$ using the method in Equations 3-47, 3-49, and 3-50. We will construct the observed measures using the standard EM algorithm for Gaussian mixtures.

Results are illustrated and further detailed in Figure 3-3. Note that the resulting likelihood from EM optimization results in a measure learned from the set of all singleton samples, whereas, the learned representative function is a measure which was learned from a set of measures. If the distribution of *X* is similar to that of each X_i with respect to the number of samples in the distribution, the representative measure learned will be similar to the likelihood learn using standard methods. This is because all information can be detailed without any set information; however, if the distributions are different with respect to the number of samples, the learned

measures will be different. This result is illustrated in Figure 3-3. This distinction is a direct result of the proposed methods ability treat each set as a unitary element.

We have identified some fundamental differences between the proposed method and standard techniques. Note there are some similarities and differences when performing inference using standard techniques and the proposed technique. In many cases the calculation of likelihood is different during inference; however in some cases, the result of inference—determination of the most probable context—is similar. In fact, if the representative measure q is the actual learned likelihood of the standard method, that is $p(x \mid y) = q(x) \; \forall x$, then the result of inference will be the same. This shared similarity between the two approaches is again shared if the distribution of X is similar to that in each X_i.

Example 3.7 Inference similarities and differences: The random measure approach assigns high likelihood to sets, or random measure instances, that have a similar distribution throughout the domain; whereas standard approaches assign high likelihood as long as each observed singleton sample appears in a place of high likelihood. This difference is illustrated in Figure 3-3C, which continues from Example 3.6. Note that although this is a fundamental difference, the result of context estimation may be similar using both approaches depending on the observed measures construction and the results of optimization.

During the optimization of the proposed likelihood function, the representative function is learned. This is similar to variational methods where functions are learned by optimizing objective functionals.

Example 3.8 Comparison with standard inference using variational methods: Given an observed set X, we want to determine if it was observed in context i. Using the proposed method on random measures, we would first construct ν_X. Next, we could determine the unnormalized likelihood of some context using $p(\nu_X \mid M_{\Xi_i}) = \exp\big(-KL(P_X \mid\mid Q_i)\big)$. Whereas, with a standard variational method, or most standard methods of inference given a joint observation set, the initial observation set is explicitly assumed i.i.d., during optimization and inference. For example, the standard initial assumption made in variational inference given a set X is

$$p(X \mid Z) = \left(\frac{\tau}{2\pi}\right)^{N/2} \exp\left\{-\frac{\tau}{2}\sum_{n=1}^{N}\left(x_n - \mu\right)^2\right\} \tag{3-54}$$

where $Z = \{\mu, \tau\}$ [97]. Therefore, the estimate of the posterior $p(Z|X)$ also is i.i.d.

In Example 3-9, we compared the proposed method to standard variational methods; however, we ignored the use of hyperparameters. The hyperparameters would be better suited for contextual inference since they govern distributions on distributions and inference could be performed on observed measures. We explore the viability of using the subsequent hyperparameters for a means of context estimation.

Example 3.9 Context estimation using variational methods: For the construction of the hyperparameters, assume Equation 3-54. We then model parameters μ and τ using a normal and gamma distribution, respectively.

$$p(\mu \mid \tau) = N\left(\mu \mid \mu_i, (\lambda_i \tau)^{-1}\right)$$
$$p(\tau) = Gamma\left(\tau \mid \alpha_i, \beta_i\right) \quad (3\text{-}55)$$

It can be shown that the parameter μ_i is updated using

$$\mu_i^{t+1} = \frac{\lambda_i \mu_i^t + N\bar{x}}{\lambda_i + N}. \quad (3\text{-}56)$$

We note this is similar to update Equation 3-41, save the expectation over sets used by the proposed method. Therefore it cannot treat set values as unitary elements and will differ from the proposed method similarly to standard statistical methods, as illustrated in Examples 3-6 and 3-7.

Again, note that Equation 3-56 is somewhat similar to the optimization of the random set, where the Gaussian is the measure resulting from the update Equations 3-41, 3-42, and 3-43. The difference here is that there is a prior distribution on the parameters of some family of distributions. This simplifies computation to some degree as the random element is reduced to being a standard random variable in $\mathcal{R}^d$. Note that this is an *atypical use of the intermediate constructs of standard variational inference;* however, this potential use fits the problem of developing a likelihood on functions given a simple model.

Example 3.10 Context estimation using a mixture of Gaussian hyperparameters: In Example 3.9, we constructed a hyperparameter given a single Gaussian measure constructed from an observation set *X*. We can similarly construct a mixture of Gaussians given an observation set *X*. Given a set of observed parameters $\{\boldsymbol{\mu}, \boldsymbol{\Lambda}\}$, developed from some observed set *X*, we can estimate the likelihood of some context *i* given some trained parameters $\{\mathbf{m}_i, \boldsymbol{\beta}_i, \mathbf{W}_i, \mathbf{V}_i\}$, learned given an assumed Gaussian-Wishart prior governing the mean and precision of each Gaussian component

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda} \mid \mathbf{m}_i, \boldsymbol{\beta}_i, \mathbf{W}_i, \mathbf{V}_i) = \prod_{j=1}^{J} N\left(\mu_j \mid m_{ij}, (\beta_{ij}\Lambda_j^{-1})\right) W\left(\Lambda_j \mid W_{ij}, V_{ij}\right). \quad (3\text{-}57)$$

This development by Bishop [89] has surfaced a few inherent issues that accompany this approach. First there is the assumption that the hyperparameters is factorizable, which was previously mentioned and may or may not be that constraining dependent on application area. However, the fact that we are now performing inference in the parameter space, rather than the space of measures has lead to other issues. Note that μ_j and m_{ij} are both indexed by *j*, although they both are elements of sets $\boldsymbol{\mu}$ and $\mathbf{m}_i$, respectively. This implies that in order to properly calculate the likelihood, there must be the same number of observed samples as there are Gaussian-Wishart priors and that the observations and distribution components must be matched.

These issues are a direct result of the hyperparameters being intermediate constructs. These constructs have one purpose, which is to model one set of observations. In fact, they are not meant to be used directly since their only use is to integrate out intermediate parameters. That is why these standard variational learning should not be used for context estimation.

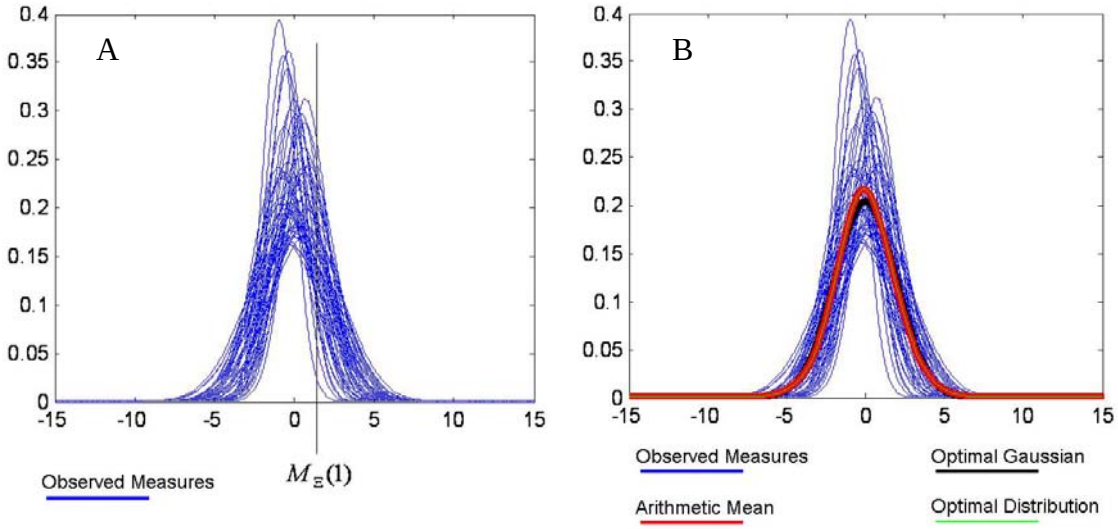

Figure 3-1. Samples of Gaussian distributions drawn using randomly selected means and variances which where drawn uniformly from a specified interval. A) Fifty sample measures are plotted. The resulting value at each point x, is a random variable. For example, random variable $M_{\Xi}(1)$ has corresponding samples that that lie on the line x=1. B) The arithmetic mean, optimal Gaussian and optimal Distribution are shown given the 50 Gaussian samples. The corresponding KL divergence values are 88.5, 91.2 and 88.5, respectively. The arithmetic mean is the optimal distribution; they coincide.

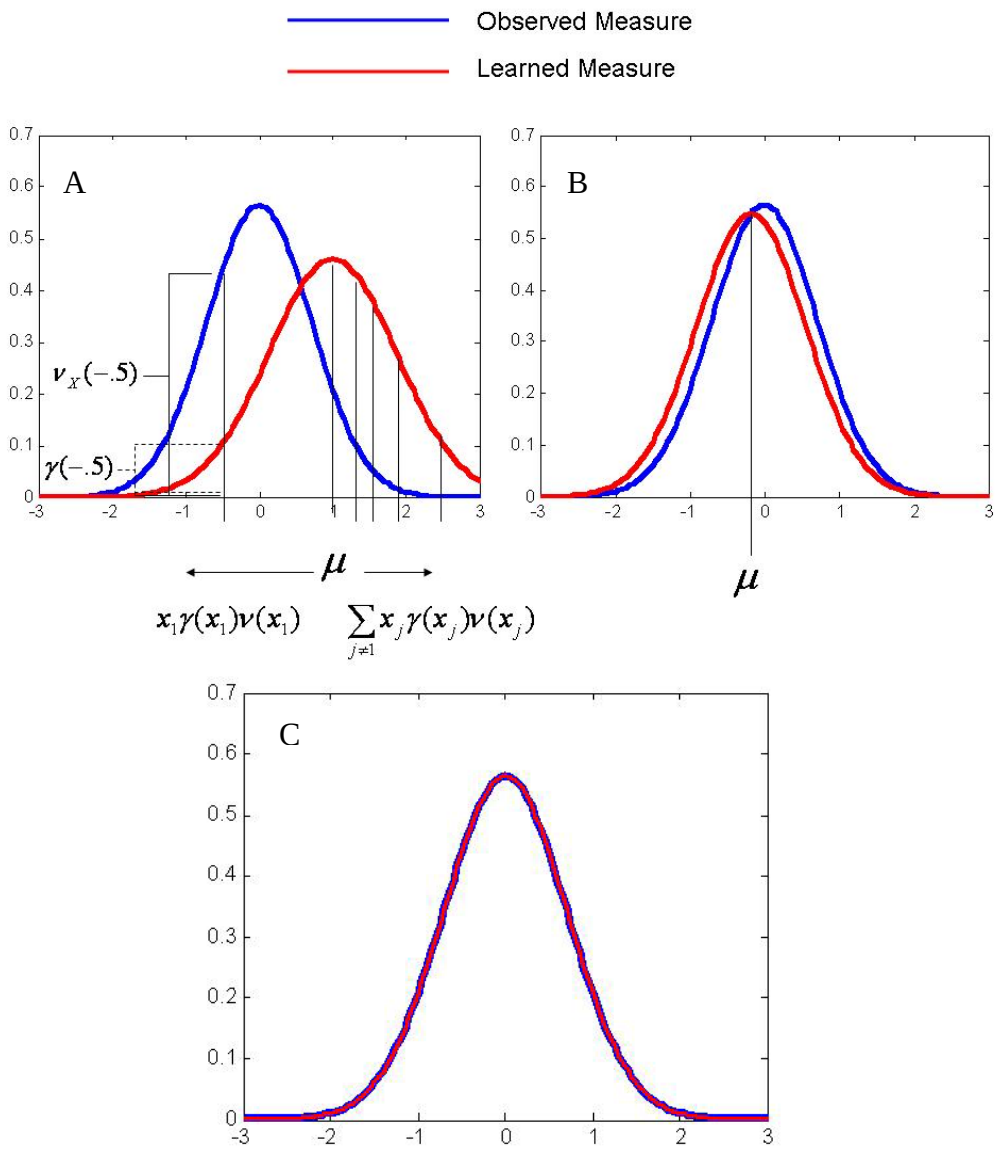

Figure 3-2. Learning the representative function using update Equations 3-47, 3-49 and 3-50 given set $A = \{-.5,1,1.3,1.5,1.9,2.5\}$. These plots illustrate the fact that the proposed method learns the function ν_X and does not fit the learned parameters to the individual samples in *A*. A) The observed measure and the initialized learned measure *q*. In standard learning techniques, optimization of the parameter μ would occur when it was the mean value of the samples in *A*, 1.28. However, the proposed objective is optimized when the correct function is learned. Parameter μ is coerced toward point $x_1 = -.5$, since $\nu_X(-.5)$ is large compared to the other samples in *A*. B) After a couple of iterations, μ becomes -.33. It should be clear that optimization coincides with function matching rather than fitting the function to the samples in *A*. C) If we use the set *A* which is a uniform sampling of 61 points in the range [-3,3], we get a better estimate of the KL divergence and the learned measure coincides with the observed measure.

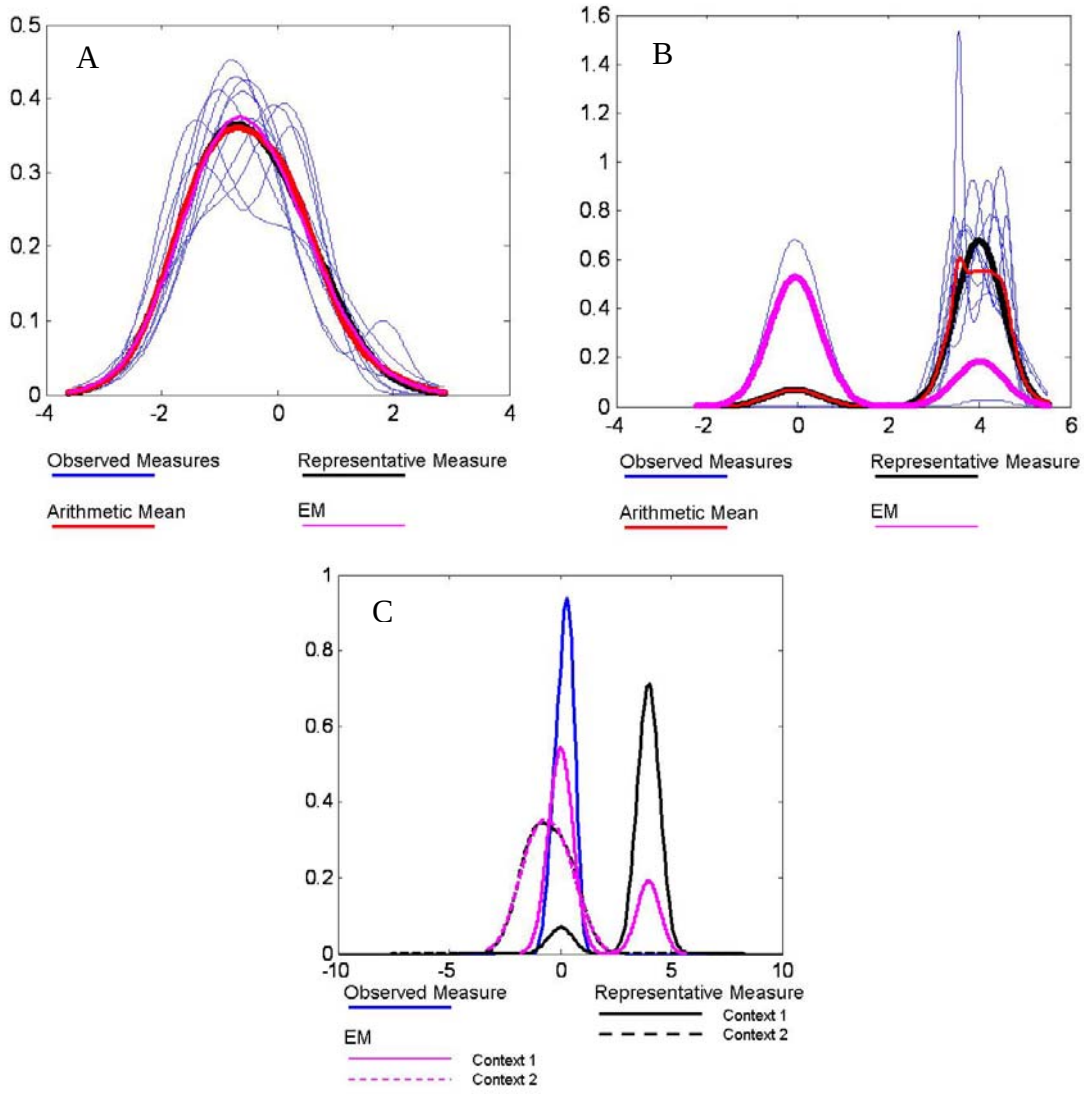

Figure 3-3. Similarities and distinctions between the proposed method and standard methods. A) The resulting EM likelihood and representative measure when optimized with respect to 10 observed sets (observed in context 2) each with a similar distribution as their union. B) The learned EM likelihood and representative measure when presented with 10 observed sets (observed in context 1) where one set has a distinct distribution compared to the union. The proposed method assumes each measure is a single sample and does not weight the one set with a different distribution any differently than the other measures. However, the standard method looks at the distribution of the singleton samples. We have constructed the set with a different distribution to have a large number of singleton samples (comparatively), to emphasize this ideological difference. C) When presented with a test set, the contextual estimates vary greatly between the standard approach and the proposed approach. Using the standard approach, context 1 is the most probable (100% to 0%); whereas, using the proposed random measure approach context 2 is the most probable (83% to 17%). Using standard i.i.d. joint estimation, the likelihood of samples lying under the

observed measure will have a greater likelihood in context 1 since the likelihood estimate for context 1 has greater likelihood values (as opposed to the likelihood for context 2) in the corresponding domain (approximately [-1, 1]). However, when comparing the representative measures for each context to the observed measure, the representative for context 2 is more similar to the observed measure.

CHAPTER 4
EXPERIMENTAL RESULTS

The three methods for context estimation developed within the random set framework (RSF) were tested using synthetic and real datasets. Four major experiments were performed. In the first experiment, we analyzed the use of different KL divergence approximation methods for estimating context in the proposed random measure model. In the second experiment, each of the three methods was tested using synthetic datasets created to imitate data in the presence of contextual factors. We compared synthetic data classification results of the proposed RSF approaches to that of set-based kNN [12], [15] and the whitening / dewhitening transform [65]. The main purpose of the experiments using synthetic data sets is to identify situational pros and cons of each of the approaches. Each method's ability to identify correct context is evaluated through its classification results since the ultimate goal is classification. Hence, we may refer to our results as context estimation results, but show the classification error on a sample basis. In the third experiment, the proposed methods are applied to an extensive hyperspectral data set collected by AHI for the purposes of landmine detection. This data set exhibits the effects of contextual factors. The purpose of the experiments using real data sets is to show the applicability of the proposed random set framework to real-world problems. We compared the hyperspectral data classification results to that of set-based kNN and the whitening / dewhitening transform. In the final experiment, the possibilistic approach is compared to a similar classifier that does not use contextual information, and it is also compared to an oracle classifier that always selects the correct context for the purposes of context-based classification. These comparisons compare the possibilistic approach to, informally, its lower and upper bounds.

KL Estimation Experiment

Experiment 1 demonstrates the efficacy of five different constructions of the set *A* for KL estimation in the proposed random measure model. Recall that if Equation 3-32 is used for KL estimation, the set *A* must be constructed such that it admits a tractable calculation but not at the expense of correctness. Therefore, we varied both the construction and size of *A* and analyzed how each affects its ability to estimate context.

Experimental Design

We compared the results of context estimation using three synthetic datasets. In the experiments each training set is constructed randomly by sampling from a Gaussian mixture with two components. Three Gaussian mixtures are used to simulate three distinct contexts. Fifteen samples are generated from each component in each context. This experimental design attempts to simulate a two class problem within each of three contexts. Ten training populations are constructed from each of the three Gaussian mixtures to simulate sets of samples observed in 3 distinct contexts. Observed measures are then created using Equation 2-40 and assuming ς is a Gaussian distribution; training is performed using Equations 3-47, 3-49, and 3-50. A test population is then generated from one of the Gaussian mixtures, which is randomly selected, and its corresponding observed measure is created assuming it is Gaussian. The representative function in the random measure model is learned from the 10 training measures and used to estimate the correct context of the test measure.

Experiments were performed using three data sets where each data set, from one to three, represents an increasingly difficult context estimate problem due to highly overlapping contexts. The data sets are Gaussian sample sets, so all experiments are repeated 50 times. Examples of each dataset are illustrated in Figure 4-1.

Each of the random measure models under test uses Equation 3-32 to estimate the KL divergence and performs contextual estimation using the random measure likelihood function as in Equation 3-32. The five methods used to construct the set *A* are as follows. In the Riemann test method, *A* is composed of N^d samples constructed by taking all combinations of test sample values in all dimensions and a Riemann integral is approximated. In Riemann test and train method, *A* is a constructed as in the Riemann test, but the samples are constructed using testing and training samples. In the naïve test method, *A* contains the observed test samples and Δx is assumed constant. In the naïve test and train method, *A* contains the observed test samples and the observed training samples and Δx is assumed constant. In the uniform MCMC method, *A* is the result of sampling a uniform distribution from within the hyperrectangle covering the train and test samples and Δx is assumed constant. Note the Riemann test and Riemann test and train methods are the same during the training phase, but differ during testing. The same is true for naïve test and naïve test and train methods. We point this out since during training only the training samples are used by all of the methods.

The Riemann approaches approximate the Riemann integral, which is a fairly standard approach. However it may be intractable for high dimensional data. Using the observed samples to partition the space into these grids would require an exponential number of elements in *A* with respect to the number of observed samples.

In the naïve test approach, *A* is simply the observed test samples and Δx is assumed constant. In the naïve test and train approach, *A* is simply the union of the test and training samples of the particular context which is to be inferred. We note these approaches are very tractable but we hypothesize that they will not be good estimates of the KL divergence.

The uniform MCMC method constructs a hyperrectangle that covers the testing and training sets using simple min and max operations. Then a fixed number of samples, the same number of samples in the observed set X in this experiment, are uniformly sampled from the covering hyperrectangle and Δx is constant for the approximation. The intuition behind this approach is that if the samples are truly uniform, Δx should be similar for each sample. The hypothesis is that this method will balance tractability and correctness.

Fifty experiments are run on each of the three data sets. For each method the representative measure is assumed to be Gaussian. The resulting contextual estimation results are compared to those attained by the random measure model using the analytical KL solution. The error of the methods under test is the average difference between themselves and the analytical solution, which is assumed to be the correct estimate. We also compare the contextual estimation error as a function of the number of observed samples. The hypothesis is that as the number of samples increases, the KL estimates will improve.

Results

The results of context estimation are shown in Table 4-1. The Riemann approaches have the least error total for all three data sets. Uniform MCMC had a low error and performed slightly better than the Riemann test and train method for datasets 2 and 3. The naïve methods had the most error for each data set, and the naïve test method had the maximum error, 8.7%, on data set 2.

Interestingly, Riemann test, which only uses the test samples for estimation purposes, performs better than Riemann test and train which uses both test and train samples. This is due to our Riemann approximation. Due to the construction of A, Riemann test and train, will have considerably more elements in the set A. Although more elements may mean higher granularity and potentially a better estimate of Δx, it has also exacerbated error in estimation. We used the

upper bound estimate to approximate the integral, which means KL estimates are slightly high for each Δx. Therefore, if we have considerably more Δx we may have more error, even with the better granularity.

Given the error estimates, the uniform MCMC seems to perform similarly to the Riemann estimates. However, it takes much less time to calculate than the Riemann approaches. Figure 4-2A shows a plot of context estimation error versus the number of samples in the initial observation set. For the Riemann approaches, there are exponentially many points that are added to correctly partition the space like a grid. On the other hand, the uniform MCMC approach performs uniform sampling and constructs *A* to have the same number of samples that are in the observation set.

Figure 4-2B shows the computation time needed for the Riemann test and train and uniform MCMC methods versus the size of the observed set. Although the Riemann approaches perform slightly better at integral estimation, uniform MCMC does comparably well and needs a very small amount of relative computation time. The runtime for the Riemann approach is exponential with respect to the number of observed samples, whereas the uniform approximation has a linear relationship as shown in Figure 4-2C.

Synthetic Data Experiment

The classification ability of the methods is under test in this experiment. Again synthetic data is created to simulate the effects of contextual factors. Four data sets are constructed such that each exposes a pro and/or con for each of the proposed methods. Each of the four data sets are illustrated in Figure 4-3 which helps to visualize the experimental setup and the purpose for each of the carefully constructed datasets. We also experiment with the whitening/dewhitening transform and set-based kNN to expose their pros and cons and for comparison purposes.

Experimental Design

Again samples are randomly generated from a Gaussian mixture with two components where samples from each component are assumed to be from the same class. Again, there are three contexts which allows for clear, less cluttered, analysis. Ten training populations are constructed from each of the three Gaussian mixtures to simulate sets of samples observed in three distinct contexts.

The contextual parameters, λ_{ij}, for the possibilistic and evidential models are optimized as described in Equations 3-23 and 3-26, respectively. In the probabilistic approach, using random measures, the observed measures are created using Equation 2-40, assuming they are Gaussian. The representative functions of the random measure likelihood functions were learned using the EM algorithm in Equations 3-47, 3-49 and 3-50, in a supervised manner. That is, each model's representative function was optimized using only the samples from the corresponding context.

We performed 50 trials on each data set; in each trial, a test set was generated randomly from one of the Gaussian mixtures associated with one of the contexts. For the random measure model, the corresponding measure was created using the standard EM algorithm assuming a Gaussian mixture of two components. The proposed evidential, probabilistic and possibilistic methods were equipped with Gaussian mixtures optimized separately using the standard EM algorithm. The contextual components were optimized separately as previously discussed.

The set-based kNN algorithm assigned, to each test sample, the label of the closest training sample in the closest set, that is, $k = 1$. The whiten / dewhiten transform was calculated as described in Equation 2-10 for each training image. The resulting confidence value was simply averaged over the training sets, since this algorithm does not provide for context estimation or relevance weighting.

Data set 1 is a fairly simple data set which should allow for simple context estimation and, within each context, simple classification. An example of data generated under data set 1 is shown in Figure 4-3. There are some disguising transformations present; however, the hypothesis is that most of the classifiers will perform well since context estimation is fairly simple in this data set.

In data set 2, we orient the Gaussians such that samples from class ‘x’ are relatively the same as compared to the samples from class ‘o’ in each of the three contexts. This data set was constructed to highlight the fact that the whitening / dewhitening transform assumes similar orientation of classes throughout each context. Therefore, the hypothesis is that the whitening/dewhitening transform will perform well on this data set. Each of the other methods should perform well since there remains only a slight presence of disguising transformations, and context estimation is therefore simple.

In data set 3, we introduce the presence of an outlier in the test set. The hypothesis is that the possibilistic approach should remain a good classifier since it has shown to be robust [16]. The evidential estimate will be affected by the outlier since it is a pessimistic approach. The probabilistic approach may be slightly affected if the observed measure is skewed toward the outlier. Set-based kNN will be affected by the outlier due to the use of the Hausdorff metric. The whitening / dewhitening transform may be affected since the outlier may drastically influence the whitening process.

In data set 4, we introduce multiple outliers which are placed relatively near to the observed samples. This data set is constructed to alter the observed measure and therefore, our hypothesis is that the probabilistic approach will be highly affected, along with the evidential approach and set-based kNN. The possibilistic approach should be unaffected by the outliers.

Classification results from the whitening / dewhitening transform will be drastically changed if the outliers greatly skew the whitening process.

Lastly, we analyzed the classification results of the evidential and possibilistic approaches, on dataset 3, when the number of germ / grain pairs was varied.

Results

The average classification errors are presented for each classifier for each dataset in Table 4-2. In data set 1, each method performed with under a 10% error and the best method, the evidential model, performed best with a 4.1% error. The whitening / dewhitening transform performed the worst since it relies on the fact that each class is relatively oriented in the same manner throughout each context, which is not the case (slightly) in data set1. The possibilistic approach performed the worst out of the proposed methods. Upon inspection, it fails to correctly identify context when an observed sample falls near a germ of an incorrect context. This is illustrated in Figure 4-3A. In the trial illustrated in Figure 4-3A, Context 3 is the most probable which is incorrectly estimated due to the close proximity of one of the test samples to the germ for context 3. The evidential and probabilistic models performed similarly, well. Set-based kNN performed slightly worse, which is attributable to the use of a nearest neighbor classifier as opposed to a Bayesian classifier.

In data set 2, the whitening / dewhitening transform results improved as expected. The evidential and probabilistic models performed equally as well. Set-based kNN and the possibilistic model performed relatively similarly.

In data set 3, the presence of an outlier drastically affected the classification results of the evidential model and set-based kNN. This data set is illustrated in Figure 4-3C. Both of the metrics used by these methods are pessimistic and are therefore affected by outliers. The

possibilistic and probabilistic approaches remained unaffected. Similarly, the whitening / dewhitening transform produced similar results as those found in data set 1.

In data set 4, the evidential model and set-based kNN remained highly affected by the presence of outliers. This data set is illustrated in Figure 4-3D. The incorporation of multiple outliers also affected the results of the probabilistic approach. The presence of multiple outliers was enough to greatly affect the observed measure and therefore tainted the context estimation. The possibilistic approach remained unaffected by the outliers. The whitening / dewhitening transform performed relatively well, although the samples from each class were not relatively oriented the same in each context. However, we note that in each context samples from class ‘o’ were to the right of samples from class ‘x’ in each class.

Table 4-3 shows the classification error, on data set 4, of the evidential and possibilistic models as the number of germ grain pairs varied. It also shows the classification error of the probabilistic approach for a baseline comparison. Overall the classification error decreases as the number of germ and grain pairs increases. This result is expected since more germ / grain pairs should allow for more detailed shape characterization.

Conversely, in standard techniques, the optimization of a statistical classifier using probability density functions may be subject to overtraining, especially if the number of densities used is increased or the number of densities is large compared to the number of training samples. In fact, if a probability density function is optimized with respect to a small number of samples the density will become focused on the few samples thus closing some abstract decision boundary tightly around said samples, causing overtraining. The overtraining during optimization corresponds to increasing the likelihood of samples in the correct probability density.

In the germ and grain model, the probability of sets of intersecting random hyperspheres increases as the hyperspheres grow. Therefore overtraining, in the aforementioned sense, is not an issue. However, optimization in the germ and grain model may cause the random radius to diverge, seemingly, the opposite of overtraining. Appropriate MCE optimization techniques, as developed here, must be implemented to prevent divergence.

However, classification error will increase with an increase in the number of germ / grain pairs, if the increase in pairs induces one of the situations outlined in Figure 4-3A, for the possibilistic approach, or Figure 4-3C and Figure 4-3D, for the evidential approach.

Hyperspectral Data Experiment

The classifiers under test were applied to remotely sensed, hyperspectral imagery collected from AHI [101], [102]. AHI was flown over an arid site at various times in the years 2002, 2003 and 2005. Eight AHI images which covered approximately 145,000m^2 were collected at altitudes of 300m and 600m with spatial resolutions of 10cm and 15cm, respectively. Each image contains 20 spectral bands after trimming and binning, ranging over LWIR wavelengths 7.88um – 9.92um. Ground truth was provided by Radzelovege *et al.* [100]. The maximum error was estimated to be less than one meter.

The scenes consisted mainly of targets, soil, dirt lanes and senescent vegetation. There are 4 types of targets. Targets of type 1 are plastic mines buried 10.2cm deep, targets of type 2 are metal mines buried 10.2cm deep, targets of type 3 are metal mines flush with the ground and targets of type 4 are circular areas of loosened soil, referred to as holes, with diameters less than one meter in length.

Since the imager was flown over the course of 4 years at various times of day, it is reasonable to assume that environmental conditions were variable. In fact, the presence of

contextual transformations including disguising transformations was confirmed, as shown in Figure 1-1.

Experimental Design

Labeled data sets were constructed from the imagery such that all samples from each data set were assumed to be observed in the same context. Training set construction was done manually since the ground truth error was large enough to prevent automation of this task. We note that the spectral signatures of all target types were similar enough to group into the same class given this data set. Each training set consisted of 10 samples from one of four classes: target, soil type 1, soil type 2, and vegetation. Therefore each training set, whose samples are assumed to be observed in the same context, consisted of 40 samples total. There were eight training populations, one from each image used to model the context of each image. Each context was modeled using four germ and grain pairs. The contextual parameters, λ_{ij}, were optimized using Equations 2-23 and 2-26 for the possibilistic and evidential approaches, respectively. Again, the probabilistic approach was trained using the EM algorithm in a supervised manner, that is, each model was optimized using only the samples from the corresponding context to be modeled. Gradient descent optimization for the evidential and possibilistic approaches was terminated after 200 iterations or sooner if the change was minimal. The germs were set to the results of *k*-means clustering of the samples of each class for each context. The learning rate for gradient descent optimization was set to 0.1.

The three context-based classifiers within the random set framework were equipped with Bayesian classifiers implemented as a mixture of Gaussians containing two components. The classification parameters were learned using the well-known EM algorithm in a supervised manner. Specifically, optimization for the mixture components modeling a particular class was

performed using only samples from the corresponding class in the corresponding context. Diagonal loading of the covariance matrices was done to mitigate the effects of low sample numbers and high dimensionality.

Set-based kNN was equipped with a simple classifier that was the inverse distance of the test sample to the closest representative of the target class, in the closest training set and k =1. This classifier permits gray level confidences, which allows for comparison to the other algorithms in the ROC curve.

The whitening/dewhitening transform was calculated as described in Equation 2-10 for each training image. The resulting confidence value was simply averaged over the training images, since this algorithm does not provide for context estimation or relevance weighting.

Test sets, or populations, were constructed from subsets of the imagery. The well-known RX algorithm [99] was run by Ed Winter from Technical Research Associates Inc. on the imagery as a pre-screener, or anomaly detector, to collect *points of interest* (POIs). There are 4,591 POIs and 1161 actual targets in the entire dataset. Sets of samples surrounding each POI in a 9x9 pixel window were collected to form test sets. This implies there is a total of 4,591 test sets each set consisting of 81 spectral signatures. Note that each test set is assumed to be a population, which means it is assumed that each sample in the set is observed in the same context. For this dataset, the 9x9 pixel window is large enough to encompass a target and background samples, but small enough to ensure that all samples have been observed in the same context.

Each sample in the test set is classified target or non-target by each of the classifiers. The probability of target is calculated for each sample within a test set and each POI is assigned a probability of target detection by taking the mean probability of target over the center samples

within a 3x3 window, since this is the standard size of a target. We note that the prescreener was not able to identify all targets in the scene, and the maximum *probability of detection* (PD) for the classifiers is 75% or 867 targets.

Cross validation is implemented at the image level, that is, spectra from a test image are not used for training purposes while said image is under test. Note that this testing procedure assumes that there exists a training population from an image other than the test image that contains samples observed in a context similar to those in the test image. We note that this may not be a valid assumption, and may make classification very difficult; however, this testing procedure mimics the testing conditions of real-world application, that is, the exact context and labels of some of the spectra from a test image may not be known *a priori*.

Classification results of all target types are presented in one *receiver operating characteristic* (ROC) curve which is shown by PD versus *false alarm rate* (FAR). We note that previous research has indicated that a minefield can be minimally detected when the PD is greater than 50% and the FAR is less than 10^{-2} FA/m^2 and is successfully detected when the PD is greater than 50% and the FAR is less than 10^{-3} FA/m^2 [100].

Results

ROC curves for each algorithm are shown in Figure 4-4. All methods performed well achieving greater than 50% PD at relatively low FARs. Note the Probabilistic RSF approach was run using the uniform sampling technique for KL estimation and using the analytical integral, assuming Gaussian. The analytical approach performed best, although it assumed Gaussian, whereas, the uniform sampling method used a Gaussian mixture with four components. Although the uniform sampling allows for a more versatile modeling scheme, the analytical calculation of the KL divergence seemed more important than versatility for correct context estimation. Due to the high dimensionality and sparsity of the data, the KL estimate using uniform sampling suffers.

ROC curves with error bars are shown in Figure 4-5. In Figure 4-5, each mine encounter is treated as a binomial distribution and the error bars illustrate a confidence window of 95%. Note the PDs are normalized to 100% for binomial estimation, and there is good separation of the possibilistic and evidential approaches with 95% confidence indicating a non-random result.

All context-based approaches performed better than the whiten / dewhiten transform save the probabilistic approach using uniform sampling. This result is expected since these approaches are able to identify relevant contexts and use this information to correctly classify samples that have undergone contextual transformations. However, the whiten / dewhiten transform performed relatively well, which indicates that some of the classification issues induced by contextual transformations can be mitigated by means of whitening the data. Figure 4-7 shows a correctly classified POI, where each context-based approach identifies a relevant context, context 3 or 4, and consequently classifies the POI correctly.

Each of the RSF classifiers performed better than the set-based kNN classifier. This is due to their ability to identify relevant contexts in a probabilistic manner rather than a nearest neighbor manner as indicated in Figure 4-10 and Figure 4-9. This is also due to the nearest neighbor approach in the classifier as indicated in Figure 4-6 and Figure 4-9. This is due to the fact that nearest neighbor approaches do not directly incorporate the idea of probability, or weights, and therefore assign confidence based on some fixed number of samples, in this case k=1. We note in previous experiments, k=1 provided the best results for set-based kNN [17].

In Figure 4-8, a POI is incorrectly classified by the possibilistic approach. In this case, the possibilistic approach was the only method to identify context 4 as a relevant context, and consequently misclassified this POI. The situation that occurred in the synthetic data experiment that is shown in Figure 4-3A has occurred, that is, a sample from the test set has come into close

proximity of a germ / grain pair modeling context 4, which has caused the possibilistic approach to choose this context as most likely as opposed to context 3. Although the possibilistic, or optimistic approach, has caused the possibilistic approach to misclassify the POI in Figure 4-8, it also allows for resiliency in the face of outliers. An instance where the possibilistic approach chose a context different from all other approaches is shown in Figure 4-10. This POI was correctly classified by the possibilistic approach and the chance of FA was lessened by all of the probabilistic approaches as they were able to identify two relevant contexts, one which provides correct classification.

The evidential approach performed best, achieving highest PDs at almost all FARs. This result is similar to that found in the synthetic data experiment, save the situation illustrated in Figure 4-3D. The evidential approach provides a good contextual model as the inclusion functional provides an intuitive model for shape characterization.

The probabilistic approach performed well in the synthetic data experiment balancing shape characterization and robustness. However, its results in high dimensional data were inconsistent. Providing enough samples using the uniform sampling method in high dimensions was not practical, and using the analytical integral provided better results. However, the Gaussian assumption limited its shape characterization which influenced its classification results.

Upper and Lower Bounding Experiment

In this experiment we compared the proposed possibilistic context-based classifier to a standard Bayesian classifier, a non-context based classification method. We also compared the results of the possibilistic classifier to results from a context-based oracle classifier that always chooses the correct context. This comparison provides an idea of an upper bound and a lower bound for the proposed method, where the standard classifier is a lower bound since it makes no

use of contextual information and the oracle classifier is the upper bound since it makes the best use of contextual information.

Experimental Design

The experimental setup was similar to that in the hyperspectral experiment. Eight training sets were constructed each representing a set of samples, both target and non-target, observed in some context. Each training set consists of samples collected from an image, where eight contexts are modeled using samples from the eight distinct images. Note that in this training set there are 20 samples from each class in each context. Also in this experiment more spectral bands are used, that is, each image contains 40 spectral bands after trimming and binning, ranging over LWIR wavelengths 7.88um - 9.92um. Again, the possibilistic classifier is equipped with a mixture of Gaussians for sample classification. The oracle uses the same classifiers as the possibilistic approach, however, it always chooses the correct Gaussian mixture.

Classification results of the standard Bayesian classifier are compared to that of the possibilistic RSF classifier. The hypothesis is that both classifiers can account for non-disguising transformations, however, a standard Bayesian classifier cannot account for disguising transformations, whereas the possibilistic classifier can.

The number of mixture components used in the standard Bayesian classifier is varied to illustrate how its ability to classify in the presence of non-disguising transformations relates to the number of mixture components. The hypothesis is that as the number of components increases, the results should improve since it will be better equipped to handle non-disguising transformations. However, regardless of the number of mixture components, the standard Bayesian classifier cannot handle disguising transformations and its results should not best those of the possibilistic classifier, assuming context estimation is performed correctly.

The possibilistic classifier was equipped with two mixture components per class per context for a total of 56 components since for each test set there were seven contexts available each with four classes each containing two mixture components. We compared the results to those of a standard Bayesian classifier with three, seven and 14 mixture components per class.

For comparison to the upper bound, the testing procedure will remain the same, except the classifier trained on the test image will be available to the classifiers during testing; therefore, cross validation is no longer being performed. The results of the possibilistic classifier will be compared to the oracle classifier. The oracle classifier is equipped with similar Gaussian mixtures as the possibilistic RSF classifier; however, it always uses the Gaussian mixture that was trained on the test image. The results of this classifier can be seen as an upper bound of the classification results within this framework. Therefore, it provides a means to assess the ability of the context estimation methodology used in the RSF classifier, namely the optimistic germ and grain model.

Results

The use of possibilistic context estimation within the RSF significantly improved classification results. Probability of detection is improved at all FARs and is improved as much as 10 percentage points. False alarm rates are decreased at all PDs and are reduced as much as 50% at PDs of $4x10^{-3}$ FAs/m^2 through $8x10^{-3}$ FAs/m^2.

Classification results of the standard Bayesian classifier became better as the number of mixture components increased. The increase of mixture components equipped the standard classifier with the ability to account for non-disguising transformations. When the number of components was less than the number of contexts, the standard classifier performed poorly. This is expected as it could not account for all of the non-disguising transformations. However, its performance improved as the number of mixture components became greater than or equal to the

number of contexts. The results also indicate that the RSF classifier was able to account for disguising transformations with an improvement in classification when compared to the standard classifier with the same number of overall mixture components.

The RSF Bayesian classifier performed similarly to the oracle RSF Bayesian classifier indicating that using the random set framework is an excellent method for context estimation. In fact, the RSF Bayesian classifier using the germ and grain model weighted the context which was chosen by the oracle as the most likely context 66% of the time, and furthermore, weighted that context as one of the two most likely contexts 86% of the time. However, we note there is room for improvement which can be noticed at low FARs.

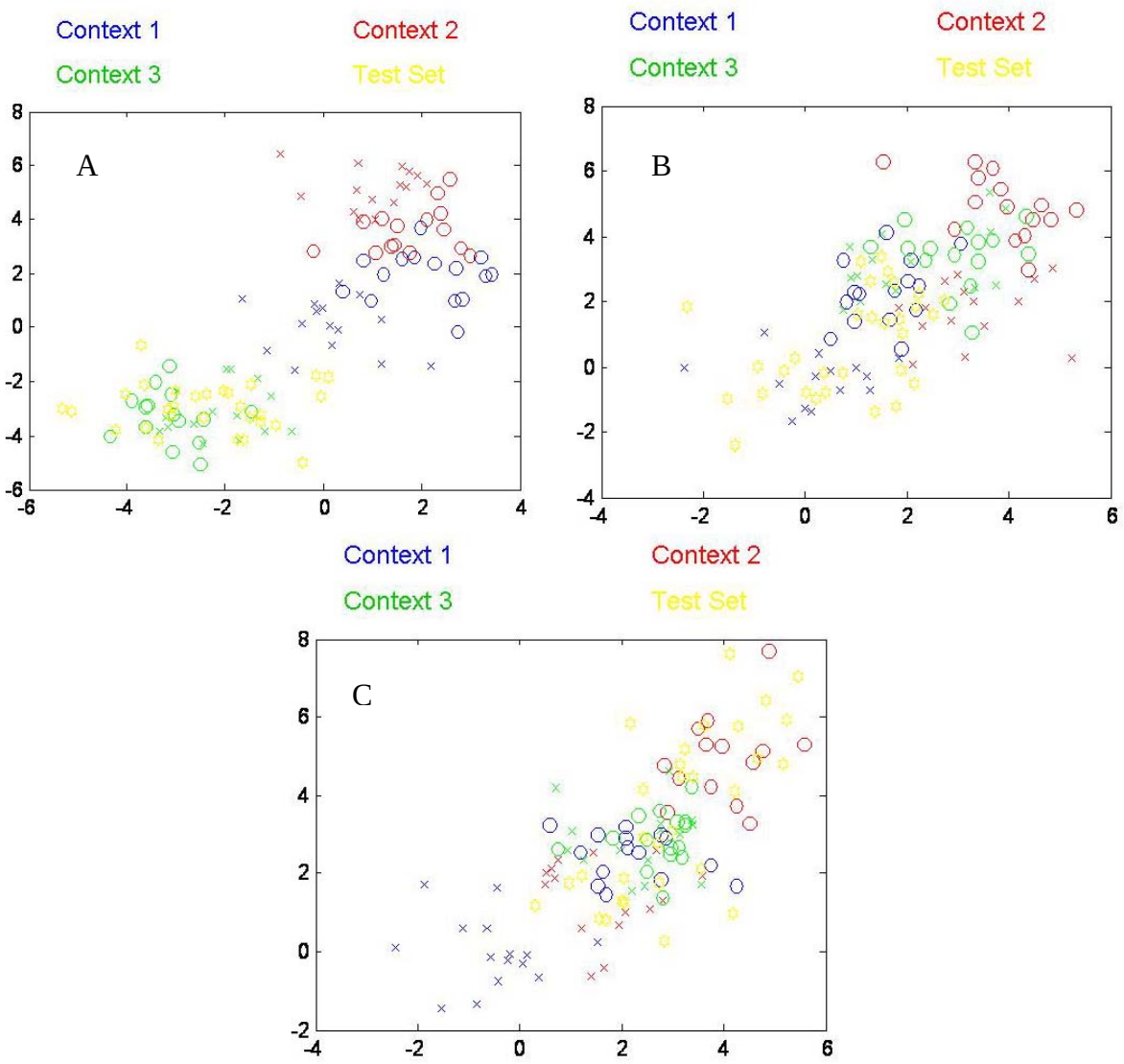

Figure 4-1. Illustration of data sets one, two, and three. A) Samples from a distinct context are shown in distinct colors. Distinct class is shown using a distinct symbol. This is the easiest data set since each context is fairly separable. B) In data set 2, context 3 is overlapped highly by both context 1 and context 2. C) In data set 3, context 1 is completely overlapped by context 2 and context 3.

Table 4-1. Average inference error for each dataset using 15 test and 15 train samples.

KL Estimation	Data Set 1	Data Set 2	Data Set 3
Riemann Test	.0094	.0390	.0522
Riemann Test and Train	.0114	.0638	.0642
Naïve Test	.0750	.0870	.0722
Naïve Test and Train	.0128	.0639	.0683
Uniform MCMC	.0094	.0562	.0581

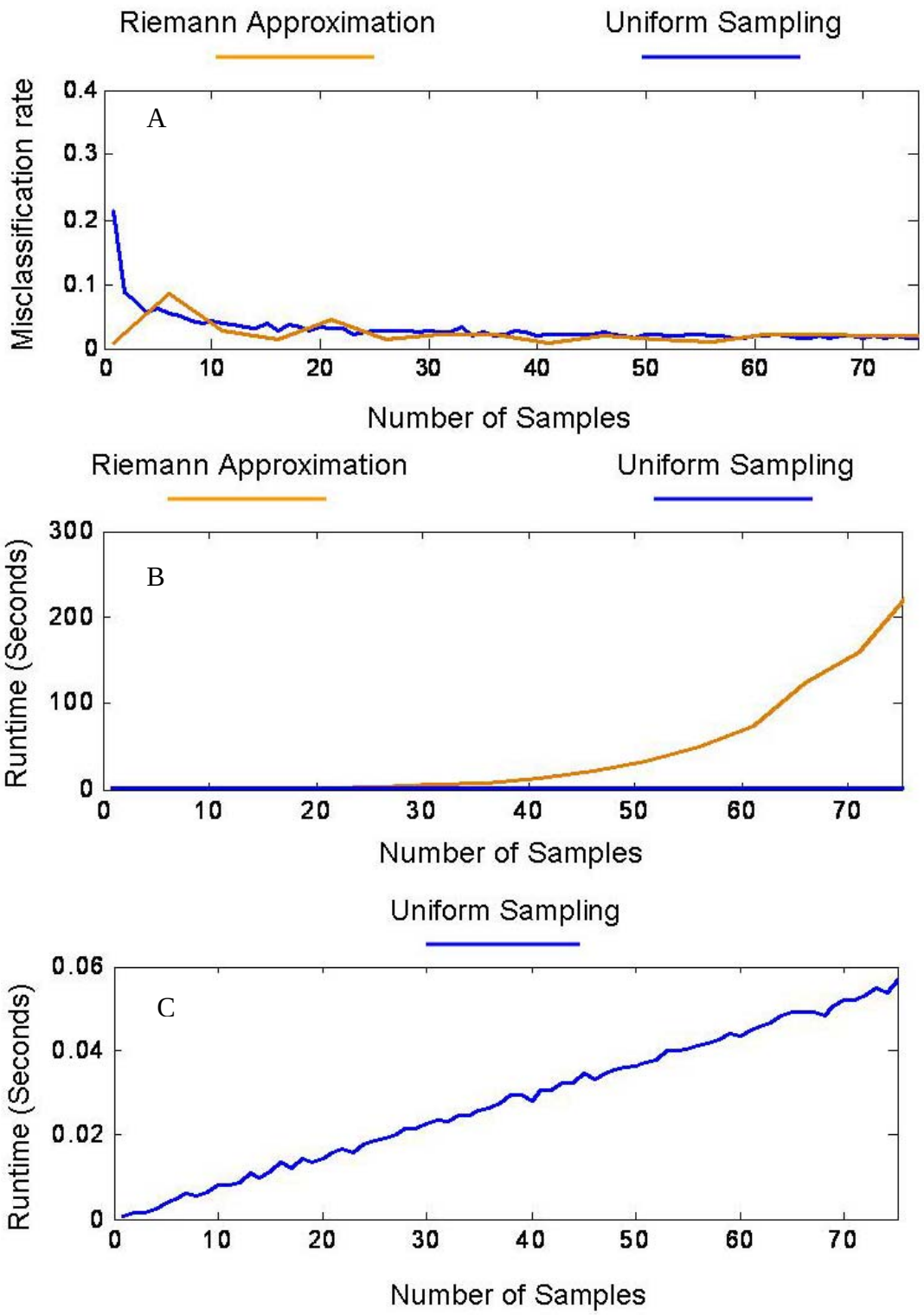

Figure 4-2. Error analysis of the Riemann and uniform approximation methods with respect to time and number of observation samples. A) Plot of context misclassification rate versus the number of samples in the observed set. B) Plot of runtime versus the number of samples in the observations set. C) Close of the plot of runtime versus number of observation samples for the uniform approximation method.

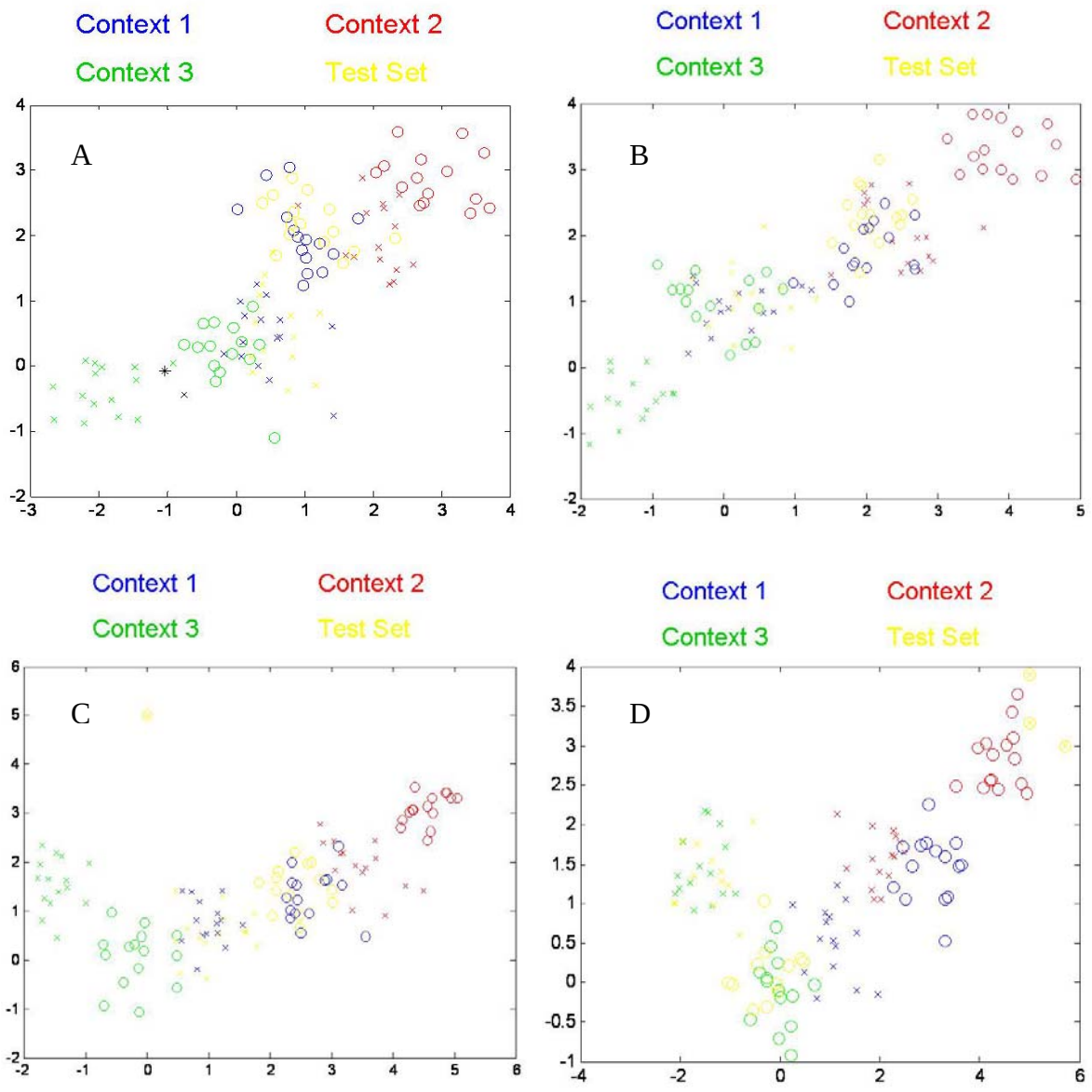

Figure 4-3. Trials using data sets 1, 2, 3 and 4 in the synthetic data experiment. A) Illustration of a trial on data set 1 from the synthetic data experiment where the possibilistic model fails to correctly identify context. Here the germ from context 3 is indicated with a black '*'. Note there is a sample from the test set indicated by a black 'x', which lies very near to the grain. This increases the probability of context 3. B) Trial example of data set 2 in the synthetic data experiment. Samples from either class are oriented relatively the same in each of the 3 contexts. C) Trial example of data set 3 in the synthetic data experiment. Each test set in each of the 50 trials has two outlying samples at [0, 5]. D) Trial example of data set 4 in the synthetic data experiment. Each test set has 6 outliers located near [5, 3.5].

Table 4-2. Average classification error of the listed context-based classifiers on four data sets used in the Synthetic Data Experiments.

Context Classifiers	Data Set 1	Data Set 2	Data Set 3	Data Set 4
Evidential Model	.0413	.0273	.2073	.2500
Probabilistic Model	.0427	.0280	.0667	.2562
Possibilistic Model	.0647	.0480	.0693	.0542
Set-Based kNN	.0560	.0373	.2647	.2520
Whiten/De-Whiten	.0993	.0220	.1033	.0792

Table 4-3. How classification varies with respect to the number of germ and grain pairs for data set 3 (with no outlying samples) in the Synthetic Data Experiment.

Context Classifiers	1 Pair/Context	2 Pair/Context	3 Pair/Context	4 Pair/Context
Evidential Model	.0447	.0487	.0367	.0373
Probabilistic Model	.0453	.0473	.0400	.0336
Possibilistic Model	.0553	.0460	.0453	.0460

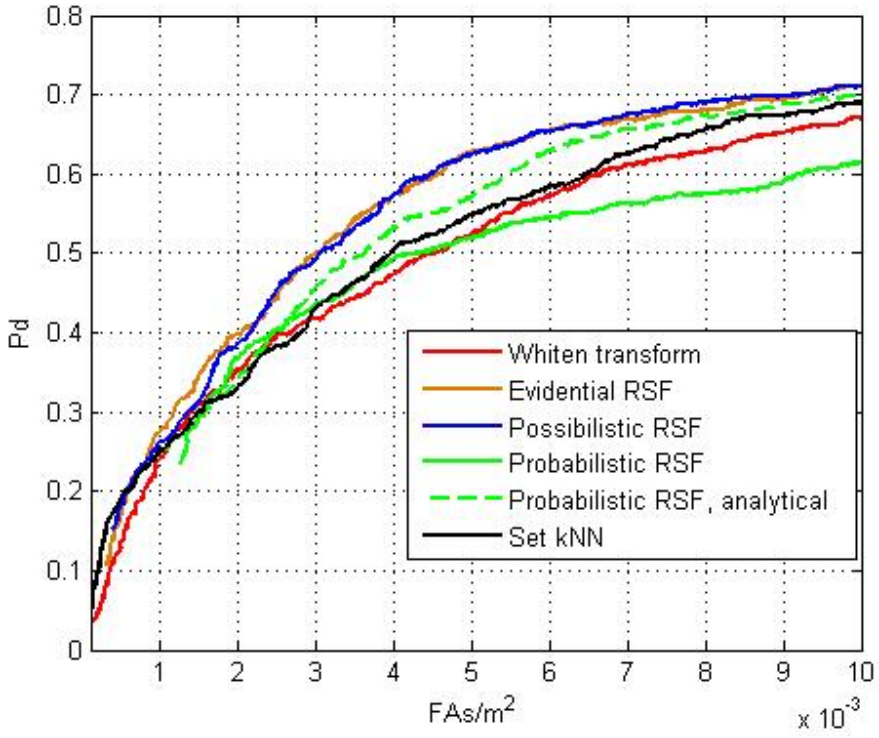

Figure 4-4. ROC curve for The Hyperspectral Data Experiment. Note the dashed plot is the results from the probabilistic context-based classifier using the analytical solution for KL estimation as discussed in Equation 3-40.

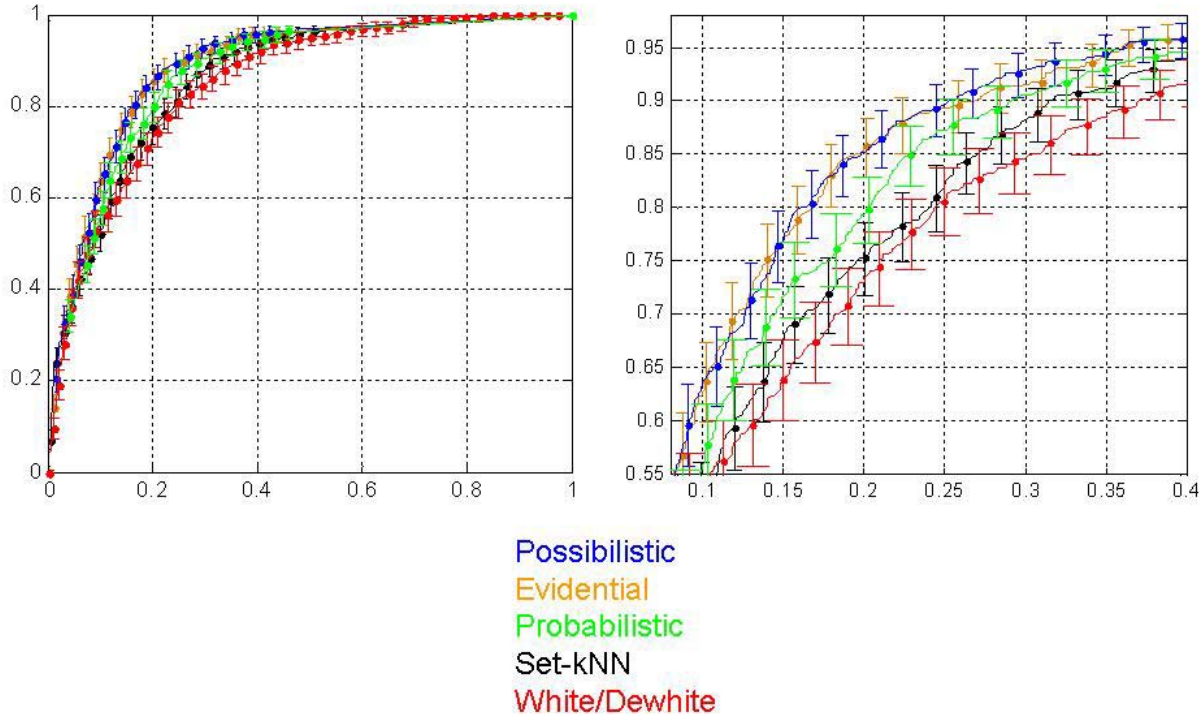

Figure 4-5. Hyperspectral Experiment ROC curve of PD versus PFA for the possibilistic, evidential probabilistic, set-based kNN, and whiten / dewhiten approaches. Error bars show the 95% confidence range assuming each encounter is a binomial experiment. For this reason, PDs are normalized to include only targets that were observed by the algorithms under test, and do not include targets missed by the prescreener.

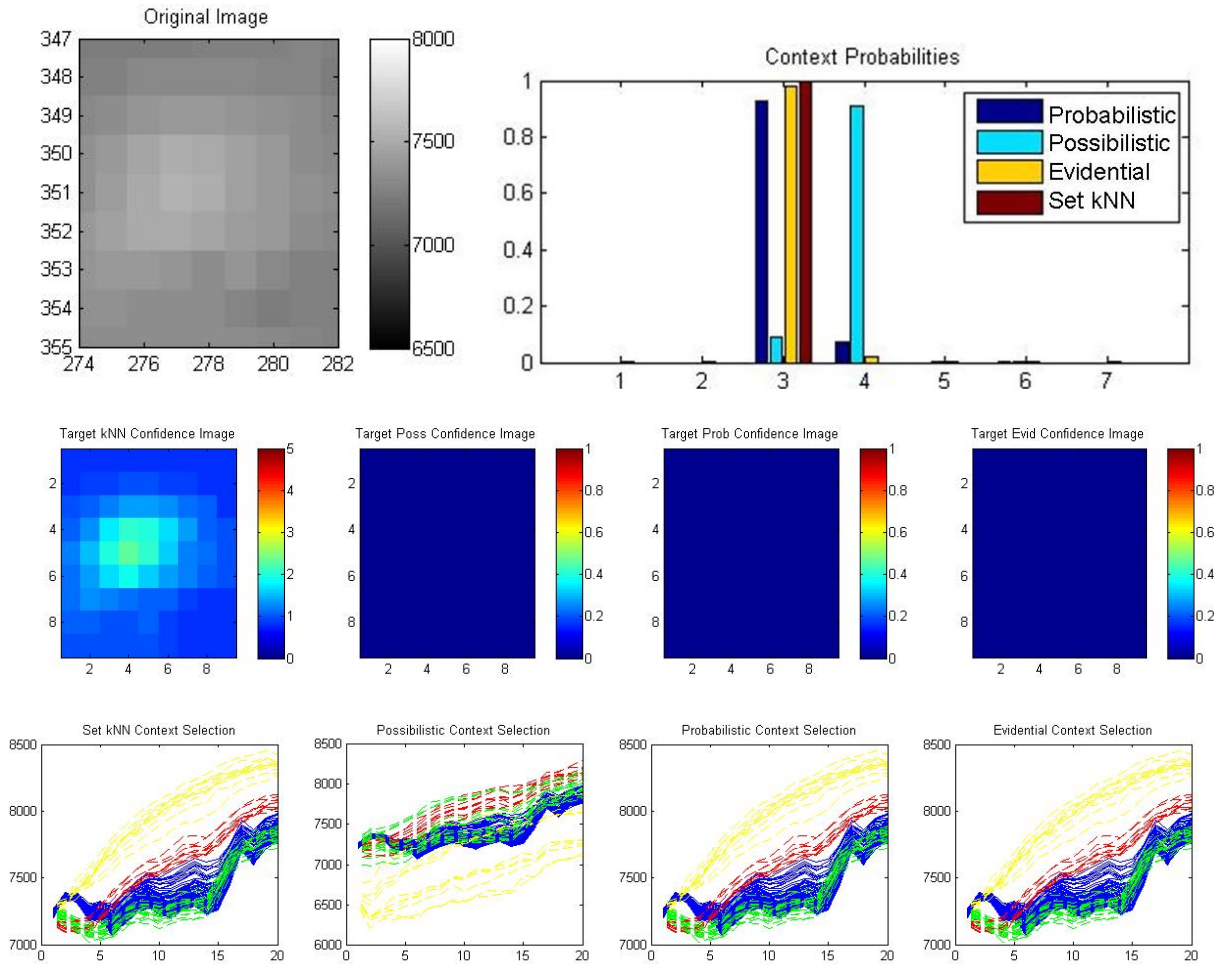

Figure 4-6. Example of a false alarm POI from The Hyperspectral Data Experiment. A snippet of the original AHI image at wavelength 8.9um is shown in the upper left where the prescreener alarmed. The second row are the confidence images of set-base kNN, possibilistic, probabilistic, and evidential approaches, from left to right. Their contextual estimates of the potential seven contexts are shown in the bar chart in the upper right. Note there are seven potential contexts and not eight since are performing crossvalidation at the image level. Under the confidence images, in the bottom row, are the spectral plots of the test population, shown in blue. Also shown in these plots are the spectra used to create the contextual models of the context that the corresponding approach selected as most probable. These training spectral are color coded by class, where red, green, and yellow correspond to target, soil types, and vegetation, respectively.

Note in this example set-based kNN submits a marginal confidence, due to its use of a nearest neighbor based classifier and choice of context 3. the probabilistic and evidential approaches select context 3 as well; but their classifier makes use of covariance which allows for correct classification. The possibilistic approach chose a context which correctly identifies the spectra as soil.

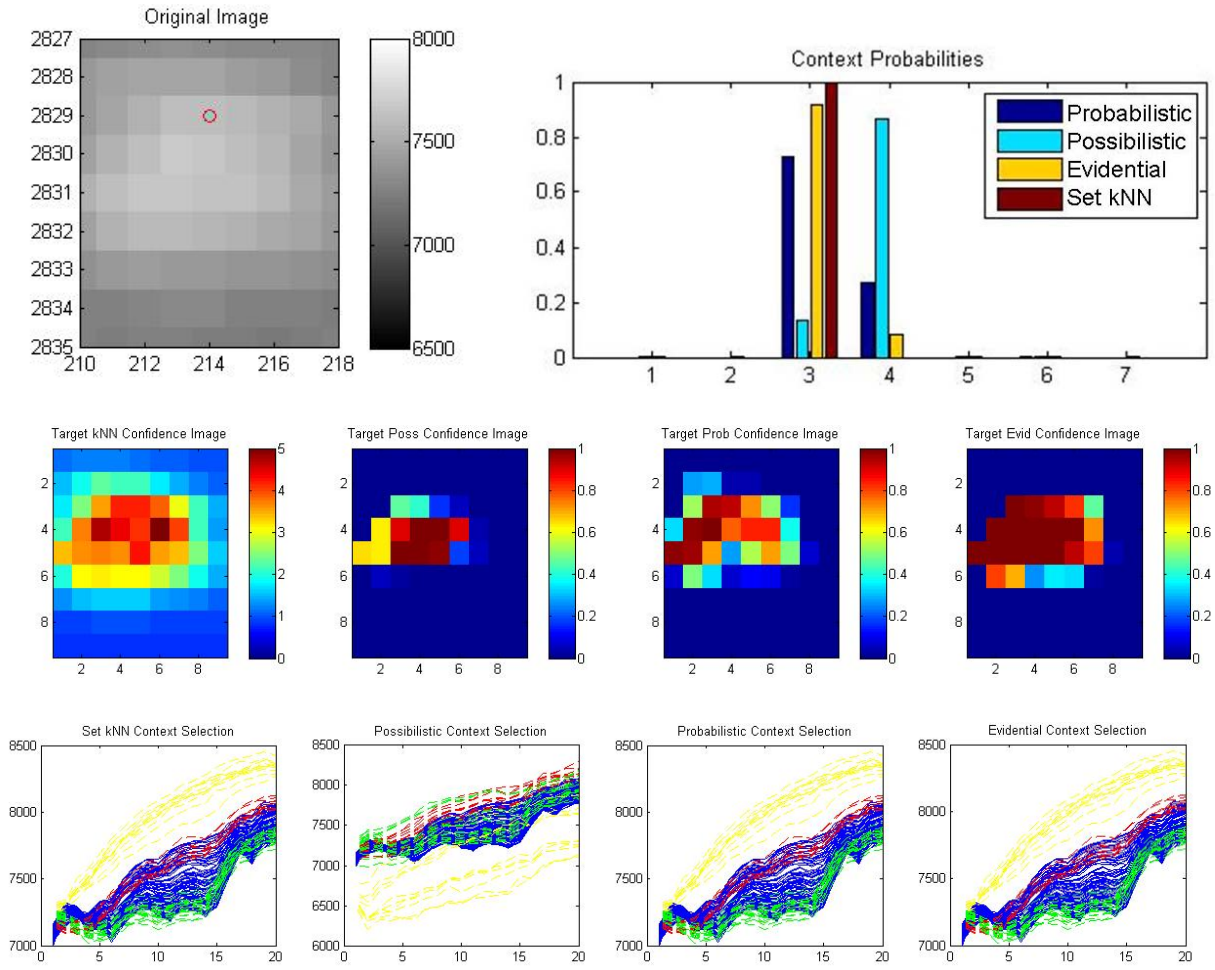

Figure 4-7. Example of a target alarm POI, from The Hyperspectral Data Experiment. A snippet of the original AHI image at wavelength 8.9um is shown in the upper left where the prescreener alarmed. The red circle indicates that this is a target. The second row are the confidence images of set-base kNN, possibilistic, probabilistic, and evidential approaches, from left to right. Their contextual estimates of the potential seven contexts are shown in the bar chart in the upper right. Under the confidence images, in the bottom row, are the spectral plots of the test population, shown in blue. Also shown in these plots are the spectra used to create the contextual models of the context that the corresponding approach selected as most probable. These training spectral are color coded by class, where red, green, and yellow correspond to target, soil types, and vegetation, respectively.

Note in this example each algorithm correctly identifies this POI as a target. Note the possibilistic approach has selected context 4 as most probable, whereas, the other three methods selected context 3. In this instance, the choice between context 3 and 4 does not change the classification results since the test spectra are similar to the target prototypes in both contexts.

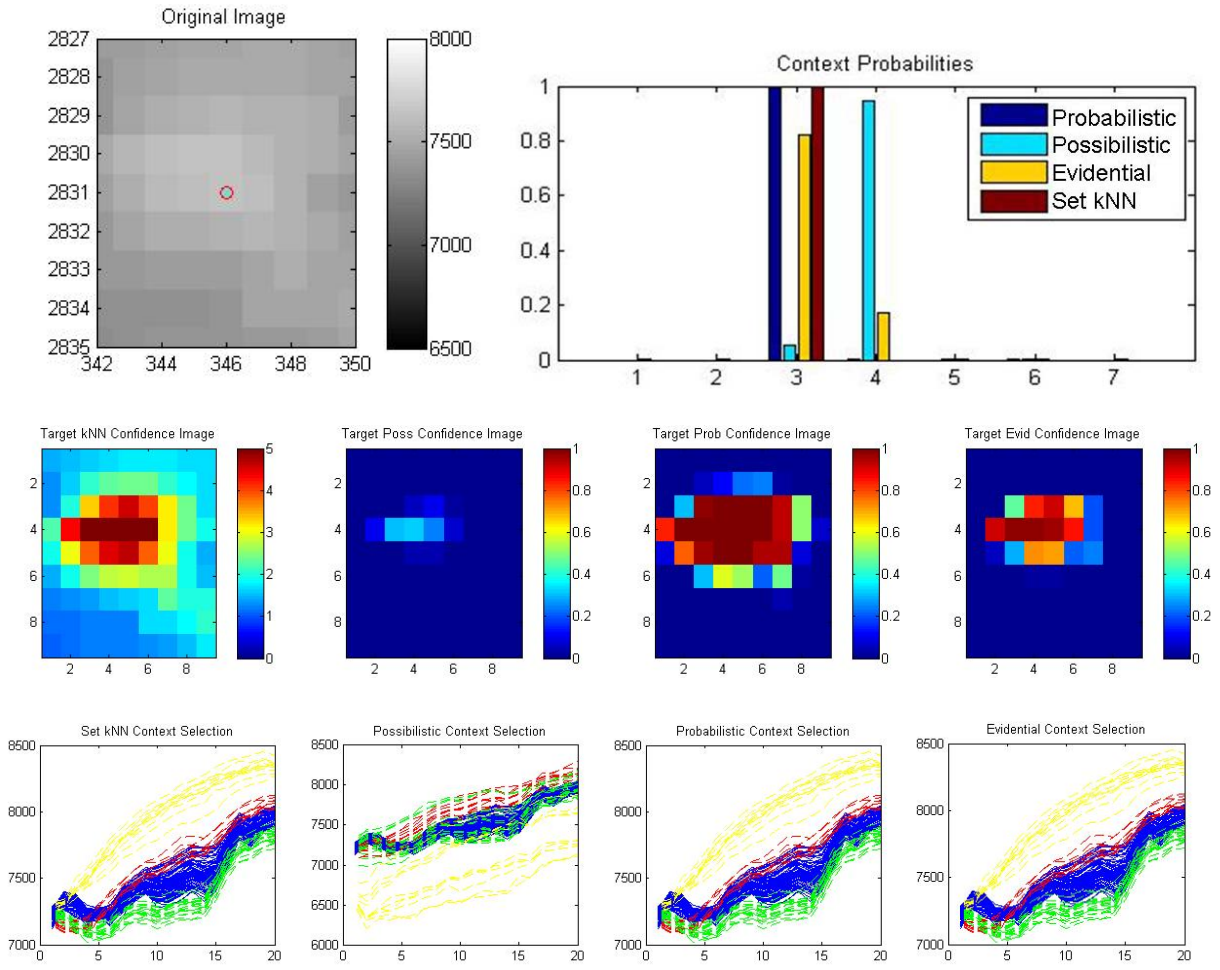

Figure 4-8. Example of a target alarm POI from The Hyperspectral Data Experiment. A snippet of the original AHI image at wavelength 8.9um is shown in the upper left where the prescreener alarmed. The red circle indicates that this is a target. The second row are the confidence images of set-base kNN, possibilistic, probabilistic, and evidential approaches, from left to right. Their contextual estimates of the potential seven contexts are shown in the bar chart in the upper right. Note there are seven potential contexts and not eight since are performing crossvalidation at the image level. Under the confidence images, in the bottom row, are the spectral plots of the test population, shown in blue. Also shown in these plots are the spectra used to create the contextual models of the context that the corresponding approach selected as most probable. These training spectral are color coded by class, where red, green, and yellow correspond to target, soil types, and vegetation, respectively.

Note in this example the possibilistic approach selects context 4, which results in incorrect classification. Also note that the evidential approach partially weights context 4, thus its confidence is not as high as set-based kNN and the probabilistic approach.

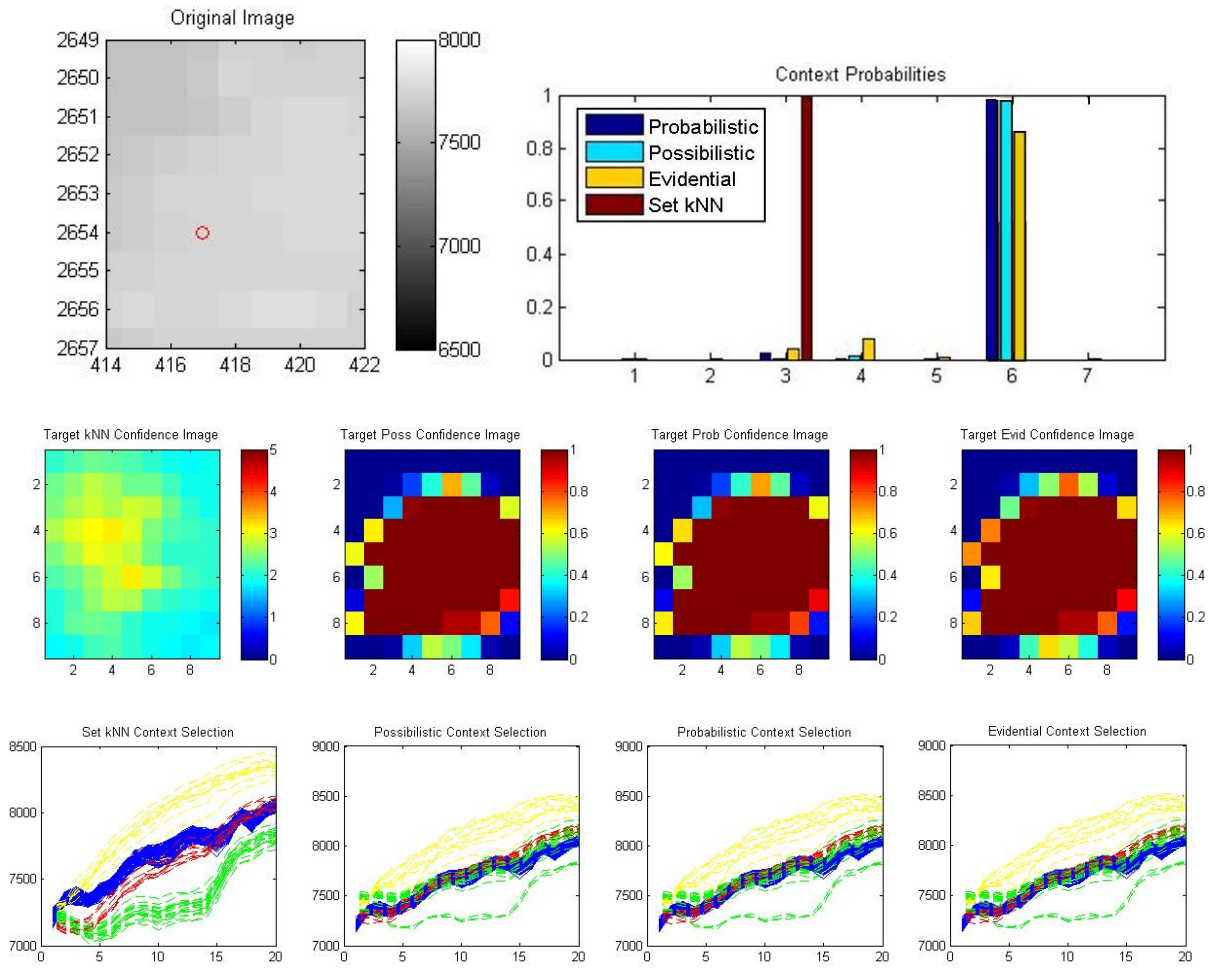

Figure 4-9. Example of a false alarm POI from The Hyperspectral Data Experiment. A snippet of the original AHI image at wavelength 8.9um is shown in the upper left where the prescreener alarmed. The red circle indicates that this is a target. The second row are the confidence images of set-base kNN, possibilistic, probabilistic, and evidential approaches, from left to right. Their contextual estimates of the potential seven contexts are shown in the bar chart in the upper right. Under the confidence images, in the bottom row, are the spectral plots of the test population, shown in blue. Also shown in these plots are the spectra used to create the contextual models of the context that the corresponding approach selected as most probable. These training spectral are color coded by class, where red, green, and yellow correspond to target, soil types, and vegetation, respectively.

Note in this example set-based kNN submits a marginal confidence rather than a high confidence due to its selection of context 3. Note the population spectra for set-based kNN selection fall in between prototypes for class target and vegetation, providing for a marginal confidence. The other 3 classifiers selected context 6 which provides for correct classification. Note samples from the target class in context 6 are extremely similar to the test samples, indicating a correct selection.

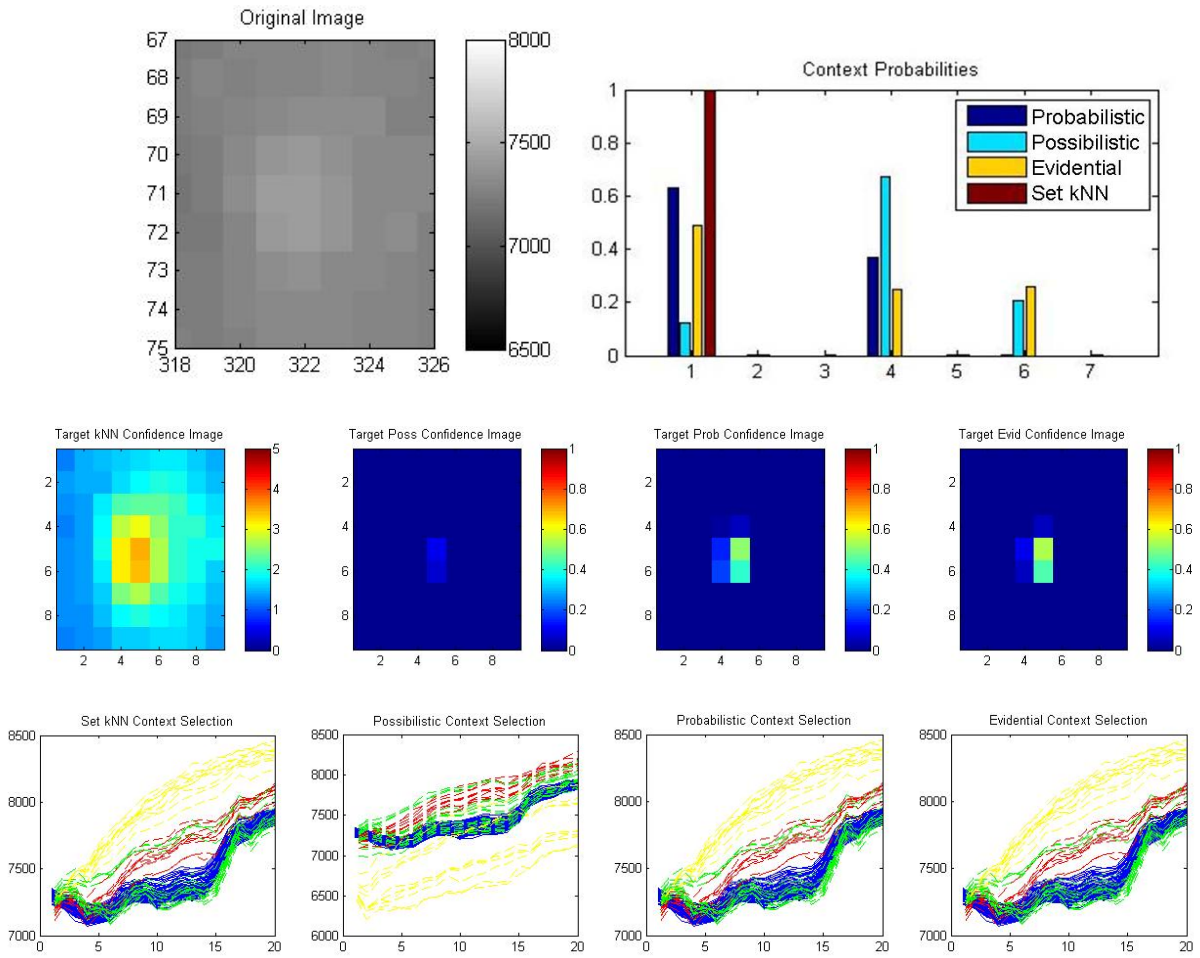

Figure 4-10. Example of a false alarm POI from The Hyperspectral Data Experiment. A snippet of the original AHI image at wavelength 8.9um is shown in the upper left where the prescreener alarmed. The second row are the confidence images of set-base kNN, possibilistic, probabilistic, and evidential approaches, from left to right. Their contextual estimates of the potential seven contexts are shown in the bar chart in the upper right. Under the confidence images, in the bottom row, are the spectral plots of the test population, shown in blue. Also shown in these plots are the spectra used to create the contextual models of the context that the corresponding approach selected as most probable. These training spectral are color coded by class, where red, green, and yellow correspond to target, soil types, and vegetation, respectively.

Note in this example set-based kNN submits high confidence due to its selection of context 1. Note the probabilistic and evidential approaches submit marginal confidences as they selected context 1. But their confidence is only marginal since they only partially selected context 1. Note the possibilistic approach selected context 4, and was able to correctly classify this POI as a false alarm.

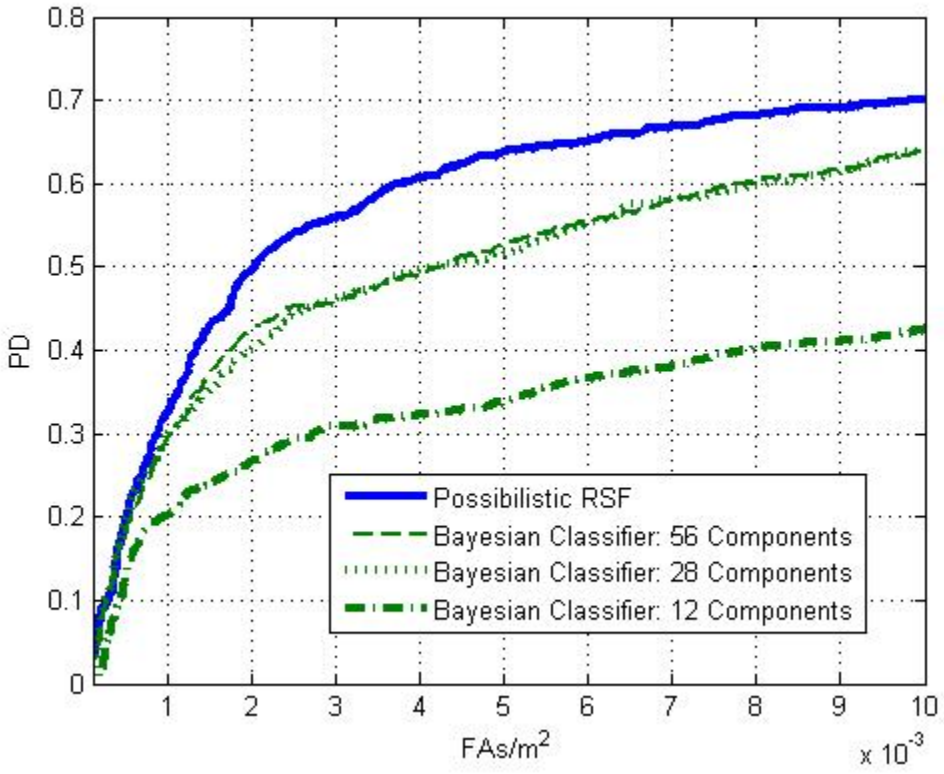

Figure 4-11. Detection results for the possibilistic RSF classifier and results for standard Gaussian mixture classifiers equipped with variable numbers of mixture components.

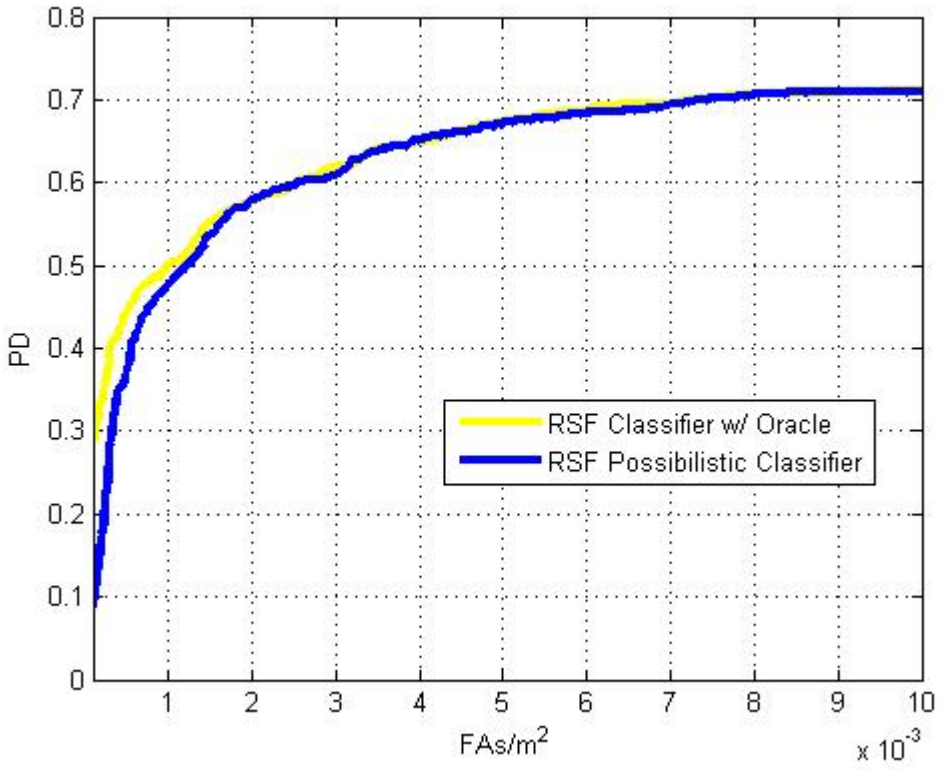

Figure 4-12. Non-crossvalidation detection results for the possibilistic RSF classifier and the oracle classifier.

CHAPTER 5
CONCLUSIONS

We developed a generalized framework for context-based classification using the theory of random sets. The resulting context-based classifier estimates the posterior of a sample, using the sample and a set—its population. Contextual transformations are identified by population analysis, and the resulting contextual estimate provides an appropriate weight of relevance to context specific classifiers. The random set framework provides the tools necessary to perform classification in the presence of contextual factors. Furthermore, it has the ability to contend with disguising transformations, which is not the case for standard classification procedures. Experimental results have shown the random set models' abilities to correctly identify context in various situations, and have shown applicability to real-world problems, improving classification results over state-of-the-art classifiers: set-based kNN and the whiten / dewhiten transform.

In the synthetic experiments, pros and cons of each approach where highlighted. The possibilistic approach was shown to be a robust classifier, resilient to outliers, but at the cost of optimism. The evidential approach has the ability to characterize shape, but at the cost of robustness. The probabilistic approach balanced these two pros and cons, allowing for some characterization of shape and some resilience. Each of these RSF classifiers was superior to set-based kNN, which is not resilient to outliers, but provides an intuitive, nearest neighbor, set comparison procedure. The whiten / dewhiten transform assumes a consistent orientation of target subspaces with respect to background subspaces, and given this assumption, provides a whitening solution. This approach can be considered a context-based method, but makes strict assumptions which the other methods do not. Therefore, the whiten / dewhiten transform performed well when said assumptions are true and performs poorly when they are not.

Each of the methods under test was able to minimally detect a minefield using a extensive hyperspectral dataset. The evidential and possibilistic methods performed best due to their resilience and shape characterization capabilities, reducing FARs by up to 25% over set-based kNN. The probabilistic model suffers partially due to its attempt to construct a representative measure given a low number samples and high dimensionality. Set-based kNN was bested by the RSF classifiers due to its lack of ability to assign gray-level weights of contextual relevance. The whiten / dewhiten transform performs worst indicating that, although some of the contextual transformations can be mitigated through the use of whitening, all of them could not.

In the final experiment, the possibilistic approach performs similarly to its upper bound, and outperformed a similar classifier that made no use of contextual information. This indicates that the possibilistic approach makes good use of contextual information which translates to improved classification results.

Each algorithm has different computational complexity. Although set-based kNN does not require training, the set-based comparison provides for a testing computation time bounded by $O(pdTN^2)$, where N is the bounding number of samples in a training or test set, d is the dimensionality of the samples, p is the number of testing populations, and T is the number of training sets. Note for each population p we must calculate the pairwise distances between the test set and all T training sets. Whereas, the RSF classifiers require a training period, but testing computation time is bounded by $O(pcdN + md^3)$, where c is determined by the fixed number of constructs, such as a germ and grain pair or a likelihood function, used to model C contexts and m is the number of constructs needed to model M classifiers. For each population, we must compare each sample to each contextual construct. Note for each Bayesian classifier we must invert a covariance matrix that is dxd; however, the use of a Gaussian classifier is not necessary

within the RSF framework. The whiten / dewhiten transform has a training period, and requires extensive testing computation time bounded by $O(pmd^3 + Nm)$. Note, for each population we must calculate and invert a covariance matrix.

Future work will include the extensive experimentation of the methods developed for the optimization and experimentation methods used by the RSF classifiers. An example of research in optimization strategies would include the investigation of the use of EM for unsupervised learning of contexts within the hyperspectral data set. This could provide for interesting findings of sub-contexts or subpopulations, within each image. Examples of future research in experimental methods would be performing experiments where the size of the populations varied. Larger populations may provide for a better estimate of context.

Extended research may include the development of a non-additive random measure. This development may provide the capability to characterize complex relationships between sets of samples, similar to a belief function. We also note that during the development of the representative function, it was determined that the point-wise average of the observed measures minimized the KL between the representative function and the observed measures, this may provide for an interesting development of posterior estimation, and relation to variational methods.

Future work should include the application of the RSF classifiers to *unexploded ordnance* (UXO) datasets. These data sets are subject to problems similar to those faced in remote sensing data, including contextual factors. The use of contextual estimation should improve classification, or target identification, for these applications as well.

LIST OF REFERENCES

[1] I. Molchanov, *Probability and Its Applications: Theory of Random Sets*. London: Springer-Verlag, 2005.

[2] J. Goutsias, R. Mahler, and H Nguyen, *Random Sets: Theory and Applications.* New York: Springer-Verlag, 1997.

[3] T. Norberg, "Convergence and Existence of Random Set Distributions" *Annals of Probability,* Vol. 12, No. 3, pp. 726-732, 1983.

[4] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic Geometry and Its Applications: Second Edition.* West Sussex: John Wiley & Sons, 1995.

[5] D. Stoyan, "Random Sets: Models and Statistics", *International Statistical Review,* Vol. 66, No. 1, pp.1-27, 1998.

[6] N. Cressie and G. Laslett, "Random Set Theory and Problems of Modeling", *SIAM Review*, Vol. 29, No. 4, pp. 557-574, December 1987.

[7] D. Hug, G. Last and W. Weil, "A Survey on Contact Distributions", *Morphology of Condensed Matter: Physics and Geometry of Spatially Complex Systems*, Springer, 2002.

[8] G. Shafer, *A Mathematical Theory of Evidence*. Princeton: Princeton University Press, 1976.

[9] M. Capinski and E. Kopp, *Measure, Integral and Probability*, New York: Springer, 1999.

[10] J. Munkres, *Topology: Second Edition.* Upper Saddle River: Prentice Hall, 2000.

[11] V. Vapnik, *Statistical Learning Theory*, New York: John Wiley & Sons, 1998.

[12] E. Dougherty and M. Brun, "A Probabilistic Theory of Clustering", *Pattern Recognition*, Vol.37, pp. 917-925, 2004.

[13] E. Dougherty, J Barrera, M. Brun, S. Kim, R. Cesar, Y. Chen, M. Bittner, and J. Trent, "Inference from Clustering with Application to Gene-Expression Microarrays", *Journal of Computational Biology*, Vol. 9, No. 1, pp. 105-126, 2002.

[14] M. Brun, C. Sima, J. Hua, J. Lowey, B. Carroll, E. Suh, and E. Dougherty, "Model-Based Evaluation of Clustering Validation Measures", *Pattern Recognition*, Vol. 40, pp. 807-824, 2007.

[15] J. Bolton and P. Gader, "Application of Set-Based Clustering to Landmine Detection with Hyperspectral Imagery", *IEEE Proceedings Geoscience and Remote Sensing*, Barcelona, July 2007.

[16] J. Bolton and P. Gader, "Random Set Model for Context-Based Classification", *IEEE World Congress on Computational Intelligence FUZZ*, Hong Kong, June 2008 (Accepted).

[17] J. Bolton and P. Gader, "Application of Context-Based Classifier to Hyperspectral Imagery for Mine Detection", *SPIE Defense and Security*, Orlando, March 2008 (Accepted).

[18] Tsymbal A, "The problem of concept drift: definitions and related work", *Technical Report TCD-CS-2004-15*, Department of Computer Science, Trinity College Dublin, Ireland, 2004.

[19] M. Bentley, "Environment and Context", *The American Journal of Psychology*, Vol. 39, No. ¼. pp. 52-61, December 1927.

[20] R. Rescoria, "Probability of Shock in the Presence and Absence of CS in Fear Conditioning", *Journal of Comparative and Physiological Psychology*, 66, pp.1-5, 1968.

[21] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Handling Local Concept Drift with Dynamic Integration of Classifiers: Domain of Antibiotic Resistance in Nosocomial Infections", *Proceedings of IEEE Symposium Computer-based Medical Systems*, 2006.

[22] M. Salganicoff, "Density Adaptive Learning and Forgetting", *Technical Report No. IRCS-93-50*, University of Pennsylvania Institute for Research in Cognitive Science, 1993.

[23] D. Widyantoro and J. Yen, "Relevant Data Expansion for Learning Concept Drift from Sparsely Labeled Data", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 3, pp. 401-412, March 2005.

[24] J. Schlimmer and R. Granger Jr., "Incremental Learning from Noisy Data", *Machine Learning*, Vol. 1, pp. 317-354, 1986.

[25] G. Widmer, "Learning in the Presence of Concept Drift and Hidden Contexts", *Machine Learning*. Vol. 23, pp. 69-101, 1996.

[26] M. Maloof and R. Michalski, "Learning Evolving Concepts Using Partial-Memory Approach", *Working Note AAAI Fall Symposium on Active Learning*, Boston, pp. 70-73, November 1995.

[27] M. Maloof and R. Michalski, "Selecting Examples for Partial Memory Learning", *Machine Learning*, Vol. 41, pp. 27-52, 2000.

[28] R. Klinkenberg, "Using Labeled and Unlabeled Data to Learn Drifting Concepts", *Workshop Notes on Learning from Temporal and Spatial Data*, Menlo Park, pp. 16-24, 2001.

[29] R. Klinkenberg and T. Joachims, "Detecting Concept Drift with Support Vector Machines", *Proceedings of the 17th Intl. Conf. on Machine Learning*, PP. 487-494, July 2000.

[30] R. Schapire, "The Strength of Weak Learnability", *Machine Learning*, Vol. 5, pp. 197-227, 1990.

[31] R. Schapire, Y. Feund, P. Bartlett, and W. Lee, "Boosting the Margin: A new Explanation for the Effectiveness of Voting Methods", *Annals of Statistics*, Vol. 26, No. 5, pp. 1651-1686.

[32] R. Schapire, "The Boosting Approach to Machine Learning: An Overview, "*MSRI Workshop on Nonlinear Estimation and Classification*", 2002.

[33] Freund, Y. and R. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*", Proceedings of Computational Learning Theory: Second European Conference*, Barcelona, 2005.

[34] E. Bauer and R. Kohavi, "An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants", *Machine Learning*, Vol. 36, pp. 105-139, 1999.

[35] E. Dura, Y. Zhang, X. Liao, G. Dobeck, and L. Carin, "Active Learning for Detection of Mine-Like Objects in Side-Scan Sonar Imagery", *IEEE Journal of Oceanic Engineering*, Vol. 30, No. 2, April 2005.

[36] Q. Liu, X. Liao, and L. Carin, "Detection of Unexploded Ordnance via Efficient Semi-Supervised and Active Learning", *IEEE Transactions Geoscience and Remote Sensing*, (submitted).

[37] Q. Lui, X. Liao, and L. Carin, "Semi-Supervised Multi Task Learning", *Neural Information and Processing Systems (NIPS)*, 2007.

[38] Y. Zhang, X. Liao, and L. Carin, "Detection of Buried Targets Via Active Selection of Labeled Data: Application to Sensing Subsurface UXO", *IEEE Transactions Geoscience and Remote Sensing*, Vol. 42, No. 11, November 2004.

[39] A. Tsymbal and S. Puuronen, "Bagging and Boosting with Dynamic Integration of Classifiers", *Proc. Principles of Data Mining and Knowledge Discovery*, PKDD, 2000.

[40] M. Skurichina and R. Duin, "Bagging, Boosting and the Random Subspace Method for Linear Classifiers", *Pattern Analysis and Applications*, Vol. 5, pp. 121-135, 2002.

[41] N. Rooney, D. Patterson, S. Anand and A. Tysmbal, "Dynamic Integration of Regression Models", *Proceedings 5th Annual Multiple Classifier Systems Workshop*, Cagliari, Italy, 2004.

[42] N. Rooney, D. Patterson, A. Tsymbal and S. Anand, "Random Subspacing for Regression Ensembles", *Proceedings 17th Intl. Florida Artificial Intelligence Research Society*, 2004.

[43] L. Breiman, "Bagging Predictors", *Machine Learning*, Vol. 24, pp. 123-140, 1996.

[44] A. Tsymbal, M. Pechenizkiy, and Padraig Cunningham, "Dynamic Integration with Random Forests", *Machine Learning*: *EMCL*, LNAI 4212, pp. 801-808, 2006.

[45] R. Schapire, "Random Forests", *Machine Learning* Vol. 45, pp. 5-32, 2001.

[46] T. Ho, "The Random Subspace Method for Constructing Decision Forests", *IEEE Transactions PAMI*, Vol. 20, No. 8, pp. 832-844.

[47] T. Ho, "Random Decision Forest", *Proceedings 3rd Intl. Conf. on Document Analysis and Recognition*, pp. 278-282, Montreal, August 1995.

[48] L. Kuncheva, "Switching Between Selection and Fusion in Combining Classifiers", *IEEE Transactions Systems, Man, Cybernetics*, Vol. 32 No. 2, April 2002.

[49] A. Santana, R. Soares, Anne Canuto, and M. Souto, "A Dynamic Classifier Selection Method to Build Ensembles Using Accuracy and Diversity", *IEEE Proc. 9th Annual Symp. On Neural Networks*, Brazil, 2006.

[50] E. Santos, R. Sabourin, and P. Maupin, "Single and Multi-Objective Genetic Algorithms for the Selection of Ensemble of Classifiers", *Proceedings Intl. Joint Conf. on Neural Networks*, Vancouver, July 2006.

[51] F. Destempes, J. Angers and M. Mignotte, "Fusion of Hidden Markov Random Field Models and Its Bayesian Estimation", *IEEE Transactions Image Processing*, Vol. 15, No. 10, October 2006.

[52] H. Frigui, L. Zhang, P. Gader, D. Ho, "Context-Dependent Fusion for Landmine Detection with Ground-Penetrating Radar", *Proceedings of SPIE*, Orlando, 2007.

[53] R. Cossu, S. Chaudhuri and L. Bruzzone, "A Context-Sensitive Bayesian Technique for the Partially Supervised Classification of Multitemporal Images", *IEEE Transactions Geoscience and Remote Sensing*, Vol. 2, No. 3, July 2005.

[54] I. Taha and J. Ghosh, "Symbolic Interpretation of Artificial Neural Networks", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 11, No. 3, May 1999.

[55] Y. Qi and R. Picard, "Context-Sensitive Bayesian Classifiers and Application to Mouse Pressure Pattern Classification" *Proceedings on IEEE Pattern Recognition*, Vol. 3, pp. 448-451, August 2002.

[56] T. Minka, "A Family of Algorithms for Approximate Bayesian Inference", Dissertation Submitted to Department of Electrical Engineering and Computer Science MIT, 2001.

[57] M. Harries and C. Sammut, "Extracting Hidden Context", *Machine Learning*, Vol. 32, pp. 101-126, 1998.

[58] M. Harries and K. Horn, "Learning Stable Concepts in Domains with Hidden Changes in Context", *Proceedings 13th ICML, Workshop on Learning in Context Sensitive Domains*, 1996.

[59] A. Berk, L. Bernstein, and D. Robertson, "MODTRAN: Moderate Resolution Model for LOWTRAN 7. Rep. AFGL-TR-83-0187, 261, [Available from Airforce Geophysical Laboratory, Hanscom Air Force Base, MA 01731], 1983.

[60] A. Berk *et. al.*, "MODTRAN4 Radiative Transfer Modeling for Atmosphere Correction", *Proceedings of SPIE Optical Spectroscopic Techniques and Instrumentation for Atmospheric and Space Research*, Vol. 3756, July 1999.

[61] J. Broadwater and R. Challappa, "Hybrid Detectors for Subpixel Targets", *IEEE Transactions Pattern Analysis and Machine Intelligence*, Vol. 29, No. 11, pp. 1891-1903, November 2007.

[62] G. Healy and D. Slater, "Models and Methods for Automated Material Identification in Hyperspectral Imagery Acquired Under Unknown Illumination and Atmospheric Conditions", IEEE Transactions Geoscience and Remote Sensing, Vol. 37, No. 6, November 1999.

[63] C. Kuan and G. Healey, "Modeling distribution changes for hyperspectral image analysis", *Optical Engineering*, Vol. 46, No. 11, 117201, pp. 1-8, November 2007.

[64] P. Fuehrer, G. Healey, B. Rauch and D. Slater, "Atmospheric Radiance Interpolation for the Modeling of Hyperspectral Data", *Proceedings. of SPIE Algorithms and Technologies ofr Multispectral, Hyperspectral, and Ultraspectral Imagery XIV*, Vol. 6966, No. 69661O-1, pp. 1-12.

[65] R. Mayer, F. Bucholtz and D. Scribner, "Object Detection by Using 'Whitening/Dewhitening' to Transform Target Signatures in Multitemporal Hyperspectral and Multispectral Imagery", *IEEE Transactions. Geoscience and Remote Sensing*, Vol. 41, No. 5, pp. 1136-1142, May 2003.

[66] S. Rajan, J. Ghosh and M. Crawford, "An Active Learning Approach to Hyperspectral Data Classification", *IEEE Transactions Geoscience and Remote Sensing*, Vol. 46, No. 4, April 2008.

[67] S. Rajan, J. Ghosh and M. Crawford, "An Active Learning Approach to Hyperspectral Data Classification"**,** *IEEE Transactions. Geoscience and Remote Sensing*, Vol. 46, No. 4, pp. 1231-1242, April 2008.

[68] R. Xu and D. Wunsch, "Survey of Clustering Algorithms", *IEEE Transactions on Neural Networks*, No. 3, Vol. 16, pp. 645-678, May 2005.

[69] H. Luo, F. Kong, K. Zhang, and L. He, "A Clustering Algorithm Based on Mathematical Morphology", *IEEE Proceedings Intelligent Control and Automation*, Dalian, pp. 6064-6067, June 2006.

[70] Y. Rubner, L. Guibas, and C. Tomasi, "A Metric for Distributions with Applications to Image Databases", *Proceedings International Conference on Computer Vision*, pp. 59-66, Bombay, 1998.

[71] Y. Rubner, L. Guibas, and C. Tomasi, "The Earth Mover's Distance, Multi-Dimensional Scaling, and Color-Based Image Retrieval", *Proceedings ARPA Image Understanding Workshop*, pp. 661-668, May 1997.

[72] H. Houissa, N. Boujemaa and H. Frigui, “Adaptive Visual Regions Category with Sets of Point of Interest”, *Lecture Notes in Computer Science*, Vol. 4261, pp. 485-493, 2006.

[73] S. Theodoridis and K. Koutroumbas, *Pattern Recognition: Second Edition*, San Diego: Elsevier, 2003.

[74] X. Descombes and J. Zerubia, “Marked Point Process in Image Analysis”, *IEEE Signal Processing Magazine*, pp. 77-84, September 2002.

[75] A. Baddeley and J. Moller, “Nearest-Neighbor Markov Point Processes and Random Sets”, *International Statistical Review*, Vol. 57, No. 2, pp. 89-121, August 1989.

[76] J. M. Billiot, J. F. Coeurjolly, and R Drouilet, “Maximum Pseudo-Likelihood Estimator for Nearest-Neighbors Gibbs Point Processes”, *arXiv:math/0601065v1*, January 2006.

[77] J. Gubner and W. B. Chang, “Nonparametric Estimation of Interaction Functions for Two-Type Pairwise Interaction Point Processes”, *Proceedings IEEE Acoustics, Speech, and Signal Processing*, Vol. 6, pp. 3981-3984, May 2001.

[78] P. Fishman and D. Snyder, “The Statistical Analysis of Space-Time Point Processes”, *IEEE Transactions Information Theory*, Vol. 22, No. 3, May 1976.

[79] R. Stoica, X. Descombes, and J. Zerubia, “Road Extraction in Remote Sensed Images Using Stochastic Geometry Framework”, *Proceedings Intl. Workshop Bayesian Inference and Maximum Entropy Methods*, France, 2000.

[80] R. Stoica, X. Descombes and J. Zerubia, “A Gibbs Point Process for Road Extraction from Remotely Sensed Images”, *Intl. Journal of Computer Vision*, Vol. 57, No. 2, pp. 121-136, 2004.

[81] X. Descombes and J. Zerubia, “Marked Point Process in Image Analysis”, *IEEE Magazine Signal Processing*, pp. 77-84, September 2002.

[82] C. Lacoste, X. Descombes and J. Zerubia, “Point Processes for Unsupervised Line Network Extraction in Remote Sensing”, *IEEE Transactions PAMI*, Vol. 27, No. 10, pp. 1568-1579, October 2005.

[83] M. Ortner, X. Descombes and J. Zerbia, “Point Processes of Segments and Rectangles for Building Extraction from Digital Elevation Models”, *Proceedings IEEE ICASSP*, 2006.

[84] L. Linnett, D. Carmichael and S. Clarke, “Texture Classification Using a Spatial Pont Process Model”, *Proceedings IEEE Vis. Image Signal Processing*, Vol. 142, No. 1, February 1995.

[85] D. Savery and G. Cloutier, “Monte Carlo Simulation of Ultrasound Backscattering by Aggregating Red Blood Cells”, *Proceedings IEEE Ultrasonics Symposium*, 2001.

[86] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions Speech Audio Process.*, Vol. 5, pp. 257–266, May 1997.

[87] S. Katagiri, B.-H. Juang, and C.-H. Lee, "Pattern recognition using a family of design algorithms based upon the generalized probabilistic descent method," *Proceedings IEEE*, Vol. 86, pp. 2345–2372, November 1998.

[88] M. G. Rahim, B.-H. Juang, and C.-H. Lee, "Discriminative utterance verification for connected digit recognition," *IEEE Transactions Speech Audio Process.*, Vol. 5, pp. 266–277, May 1997.

[89] A. Ergun, R. Barbieri, U. Eden, M. Wilson and E. Brown, "Construction of Point Processes Adaptive Filter Algorithms for Neural Systems Using Sequential Monte Carlo Methods", *IEEE Transactions Biomedical Engineering*, Vol. 54, Mo. 3, March 2007.

[90] B. Picinbono, "Time Intervals and Counting in Point Processes*"*, *IEEE Transactions Information Theory*, Vol. 50, No. 6, pp. 1336-1340, June 2004.

[91] V. Solo, "High Dimensional Point Process System Identification: PCA and Dynamic Index Models", *Proceedings of 45th IEEE Conf. Decision and Control*, San Diego, 2006.

[92] R. Sunaresan and S. Verdu, "Capacity of Queues via Point Process Channels", *IEEE Transactions Information Theory*, Vol. 52, No. 6, June 2006.

[93] J. Gubner and W. Chang, "Nonparametric Estimation of Interaction Functions for Two-Type pairwise Interaction Point Processes", *Proceedings IEEE ICASSP*, 2001.

[94] P. Diggle, T. Fiksel, P. Grabarnik, Y. Ogata, D. Stoyan and M. Tanemura, "On Parameter Estimation for Pairwise Interaction Point Processes", *International Statistical Review*, Vol. 62, No. 1, pp. 99-117, April 1994.

[95] Y. Ogata and M. Tanemura, "Likelihood Analysis of Spatial Point Processes", *Journal of the Royal Statistical Society, Series B*, Vol. 46, No. 3, pp. 496-518, 1984.

[96] V. Isham, "Multiple Markov Point Processes: Some Approximations", *Proceedings of the Royal Society of London, Series A*, Vol. 391, No.1800, pp. 39-53, January 1984.

[97] C. Bishop, *Pattern Recognition and Machine Learning*: Springer. New York, 2006.

[98] J. Chen, J. Hershey, P. Olsen and E. Yashin, "Accelerated Monte Carlo for Kullback-Leibler Divergence between Gaussian Mixture Models", *Proceedings IEEE Intl. Conference on Acoustics, Speech and Signal Processing*, pp. 4553 – 4556, 2008.

[99] X. Yu, I. S. Reed, and A. D. Stocker. "Comparative Performance Analysis of Adaptive Multispectral Detectors," *IEEE Transactions Signal Processing*, Vol. 41, No. 8, August 1993, pp. 2639-2656.

[100] W. Radzelovage and G. Maksymonko, “Lessons Learned from a Multi-Mode Infrared Airborne Minefield Detection System”, *Proceedings of the Infrared Information Symposia*, 3rd NATO-IRIS Joint Symposium, Vol. 43 No.3, July 1999, pp. 343-364.

[101] P. Lucey, T. Williams, M. Winter and E. Winter, “Two Years of Operations of AHI: an LWIR Hyperspectral Imagery”, *Proceedings SPIE*, Vol. 4030, pp. 31-40, 2003.

[102] P. Lucey, T. Williams, M. Mignard, J. Jullian, D. Kobubon, G. Allen, D. Hampton, W. Schaff, M. Schlangen, E. Winter, W. Kendall, A. Stocker, K. Horton, A. Bowman, “AHI: an Airborne Long-Wave Infrared Hyperspectral Imager”, *Proceedings SPIE*, Vol. 3431, pp. 36-43, 1998.

BIOGRAPHICAL SKETCH

Jeremy Bolton received the Bachelor of Science degree in computer engineering from the University of Florida, Gainesville, in May 2003. He received his Master of Engineering and Doctor of Philosophy from the University of Florida in December 2008.

Currently, he is a research assistant in the Computational Science and Intelligence Lab in the Computer and Information Sciences and Engineering Department at the University of Florida. Research includes the development of algorithms, methodologies, and models with a solid mathematical and/or statistical base with applications to landmine detection. Current and previous research applies these methods to a variety of data including hyperspectral, multispectral, radar, and infrared.

Jeremy Bolton is a member of IEEE Computational Intelligence Society, IEEE Geoscience and Remote Sensing Society and Society of Photographic Instrumentation Engineers.