2002

Analysis of problem representation in the application of artificial neural networks for feature classification in imagery

Michael D. Sweet
The University of Montana

ANALYSIS OF PROBLEM REPRESENTATION
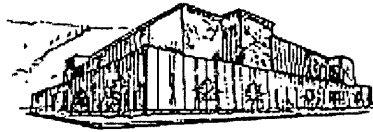
IN THE APPLICATION OF ARTIFICIAL NEURAL NETWORKS

FOR FEATURE CLASSIFICATION IN IMAGERY

by

Michael D. Sweet

B.S., Resource Conservation, University of Montana

B.A., Sociology, University of Montana

Presented in partial fulfillment of the requirements

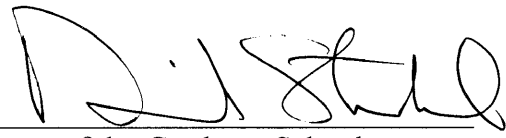for the degree of

Masters of Science
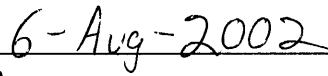
The University of Montana

August 2002

Approved by:

Chairman, Board of Examiners

Dean of the Graduate School

6 - Aug - 2002
Date

UMI Number: EP40580

UMI®

Dissertation Publishing

UMI EP40580

ProQuest®

Sweet, Michael D. M.S., August 2002   **Computer Science**

Analysis Of Problem Representation In The Application Of Artificial Neural Networks For Feature Classification In Imagery

Director: David Opitz 𝒟𝓌𝓸

The process of extracting real-world features like buildings, roads, or trees from remotely sensed digital imagery is demanding and costly. This is due to the fact that these features must represent specific concepts and characteristics in a consistent and complete manner if they are to be of value in subsequent analysis or interpretation. Machine learning approaches using artificial neural networks (ANNs) offer great promise in reducing the cost of extracting features from digital imagery.

Input representation in machine learning can affect the probability that the target concept will be learned correctly. Experiments in supervised learning examined the relationship between input representation and classification accuracy. The experiments in this study used backpropagation artificial networks and three different input representations to predict five feature classes that were both predominant and spatially distributed within a fully classified image. The experiments demonstrated that reasonable numbers of samples and learning cycles provide an acceptable probability of detecting the target when the target is spectrally distinct from other classes. In all other cases neither a large sample size, nor increased number of learning cycles, nor spatial context resulted in a significant improvement in correctly classifying a target. Inputs that were spatially clustered did not generalize as well as individual samples selected randomly from within the image. Classification accuracy was independent of proximity to training examples.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

1. INTRODUCTION

The process of extracting real-world features like buildings, roads, or trees from remotely sensed digital imagery is demanding and costly. This is due to the fact that these features must represent specific concepts and characteristics in a consistent and complete manner if they are to be of value in subsequent analysis or interpretation. Machine learning approaches using artificial neural networks (ANNs) offer great promise in reducing the cost of extracting features from digital imagery (Kanellopoulos, *et al.* 1997). They easily adapt to many problem domains and are capable of describing great complexity (Mitchell 1996). The challenge is to develop a common understanding as to how machine learning can effectively contribute to the process of classification and feature identification in digital imagery.

Marr and Poggio (1979) presented a Theory of Vision that defined perception as a series of successive approximations or varying levels of abstraction. Optimizing the learning of primitive abstractions through the classification of raw visual data may provide the necessary foundation for learning higher-level abstractions. Maximizing potential at the most primitive levels of abstraction requires evaluating the relationship between alternative input abstractions and classification accuracy within a given learning environment. This thesis examines that relationship in an effort to better understand how much of feature classification can be learned from raw visual data.

1

In this study, baseline experiments were conducted to evaluate the effect of sample size and learning cycles on classification accuracy. The experiments defined raw image data and the image pixel as the most primitive representations. This study then evaluated the effect of a higher-order input representation on classification accuracy with the expectation that it would capture information on local context that might contribute to a more accurate classification (Bain 2000). Finally, this study evaluated the spatial clustering of input pixels. It was expected that clustering pixels would be more cost effective than the random selection approach used in the baseline experiments. Clustering input examples presumes that the spatial proximity of input examples would not negatively impact classification accuracy.

The process of and errors in classification represent costs that the expert will need to assess. Results in this study confirm that primitive abstractions using pixel-level representations as inputs to artificial neural networks are cost-effective classifiers, and offer potential in advancing an understanding of the importance of primitive abstractions for feature classification. These results provide the expert with a means to assess the relative tradeoffs in the costs of classification.

2. BACKGROUND

Machine learning is concerned with computer programs that learn through experience (Mitchell 1996). Inductive learning is a discipline within machine learning, and artificial networks (ANNs) are one type of inductive learner. Like other inductive learners, ANNs learn from examples and generate a hypothesis that can be applied to unseen examples. The following sections introduce some of the concepts and presumptions of the ANN learning environment that are pertinent to experiments in this study.

2.1 Artificial Neural Networks

The most common type of ANN has an network architecture consisting of an input layer, one or more hidden layers, and an output layer, all of which are interconnected (see Figure 1 for an example). The ANN algorithm "learns" by adjusting weights on the connections. This adjustment occurs with the presentation of each training example. The output layer provides the resulting classification and has one node for each target concept to be learned. ANNs can be structured to learn a single concept (binary classification), or multiple concepts (n-ary classification). A complex network, one with more hidden nodes or hidden layers, is generally more able to achieve an accurate characterization of the training examples, but may have a lower capacity to generalize to unseen inputs. Since the problem domain generally fixes the numbers of inputs and outputs, the effective potential of simple ANNs lies in the structure of the hidden layer. An investigation by

Foody and Arora (1997) into the performance of neural network classifiers revealed that

network architecture was not as significant as other factors.
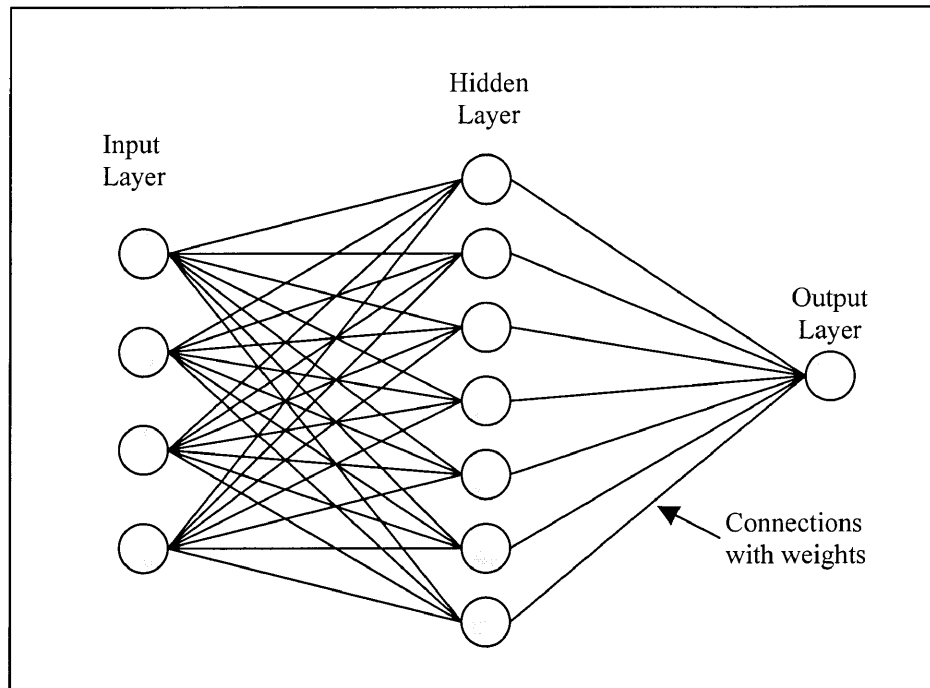
Figure 1: An example of a neural network with one hidden layer

2.2 Learning Rule

Backpropagation is a common learning rule used with ANNs to apportion error and

adjust connection weights throughout the network (Mitchell 1996). As each example is

presented to the learner, adjustments are made to the weight values. Gradient descent is

used to adjust weights to minimize error on the training set (Russell and Novig 1995). An

epoch consists of a single pass through the training set followed by verification. The

number of epochs is an ANN parameter and defines how long learning (training and

verification) continues. Too many epochs or too few can result in a model that does not generalize well. With backpropagation, as learning progresses the hidden nodes become less correlated and more independent in behavior. Wilkinson (1997) identified the performance of the learning algorithm as a key open question in the application of ANNs to image classification.

2.3 Training Sets and Test Sets

The ANN learns the target concept from a collection of examples called a training set. Another collection of examples from the image is used as a test set to evaluate the performance of the learner. There is a functional relationship between the size of the training set presented to the learner and the ability of the learner to generalize. Training and test sets should be drawn randomly, both from the same population of examples and using the same probability distribution (Russell and Novig 1995).

An investigation by Foody and Arora (1997) into the effect of training set characteristics on the performance of neural network classifiers revealed three factors having a significant effect on classification accuracy: the size of the training set; the number of discriminating variables; and the size of the testing set used for evaluation. Classification accuracy increased with the size of the training set, but the rate of increase decreased. While Foody and Arora identified significant factors, no rules were established to determine the size of the sample or the number of discriminating variables given a new

5

problem domain. The presumption is that increasing the sample size and number of learning cycles will result in a statistically significant reduction in the average error of prediction. The question remains, what is an effective sample size?

Wilkinson (1997) identified the quality of input representation as another key open question in the application of ANNs to image classification. To optimize the network's success rate the domain expert must consider how to best represent domain knowledge as inputs. Domain knowledge can focus attention on relevant inputs, or employed to define a sampling method for positive and negative training examples. There is always the assumption that input examples are correctly classified within a tolerance acceptable to the domain objective.

2.4 Spatial Context

Context consists of both local and global abstractions that are important to representing the target concept and thus provide a broader frame of reference for the learner. Context includes both spatial and non-spatial characteristics associated with a target concept, and extend to include any dependencies between components. Both are potentially important to representation and classification. In a study of the application of neural networks, Ripley (1996) presents a thorough discussion on the importance of abstraction and classification to pattern recognition. Wilkinson (1997) identified spatial context as another key open question in the application of ANNs to image classification.

Capturing information on image characteristics that are within close spatial proximity to the target concept is one means to abstract local context. Bain (2000) and Mangrich (2001) noted that foveal representations, a type of convolution filter that reduces variation by averaging surrounding pixels, captured more local context. This type of representation adds inputs to the neural network that describe pixels directly adjacent to the primary input pixel. The presumption is that changing the input representation from a single pixel to a foveal representation will capture contextual information. This in turn will result in a statistically significant reduction in classification error with less variance, while retaining a high-level of generalization. Research is inconclusive as to whether foveal representations improve classification accuracy over pixel-level representations.

2.5 Spatial Proximity

When applying ANNs to image classification an expert introduces a subjective bias through the selection of input examples and the assignment of target concepts. Subjective bias can introduce unforeseen dependencies into the classification process. One potential bias is the spatial proximity of examples, which could be even more pronounced if the domain expert rather than machine learning environment is determining the location of examples. The process of assembling input examples to the ANN is an example of where this bias may occur. Even if the location of each sample point is randomly selected, concentrating or clustering the selection of examples around the sample point may introduce a spatial bias. The presumption is that classification accuracy is a function of proximity to training examples, while examples randomly

selected throughout the range of the image will result in a predictive model with a higher capacity to generalize. In machine learning the concept of predictive error as a function of spatial proximity to the input representation is called concept-drift. There may be advantages to minimize or guide the subjective selection of input examples, and these options need to be explored.

2.6 Human-Computer Interaction

The learning assumption is that the learner must be flexible enough to capture domain specific dependencies as defined by the expert. If learning is to be efficient, an active learning environment must adapt to the stipulations of a specific problem domain and effectively link input features to target concepts to develop a representative training set. In image classification the typical target concept to be learned is membership in a class. An expert interactively assigns a target value for the selected input feature to indicate membership in the target class. This value could simply indicate presence or absence (0 or 1), or represent probability of inclusion in the target class (a value between 0 and 1). Defining the target concept as a probability of inclusion could be important in subsequent interpretations. It is unclear which membership representation of the target concept results in a better classifier, but assigning a probability of inclusion for each example would certainly be costly. In a supervised classification, deriving a valid outcome that generalizes well is contingent on consistency in the relationship between an input pattern and its assigned target classification.

2.7 Cost of Classification

Turney (1990) states that the best curve for a given data is the curve that best balances the conflicting demands of simplicity and accuracy. He suggests that simplicity in inductive inference should be defined as stability, because stability leads to repeatable experiments. Accuracy is defined as "the probability of a given data, according to the hypothesis." The tradeoffs between simplicity and accuracy are all reflected in an evaluation of the cost of learning. Even when presented with a robust set of examples, some learning algorithms are more effective than others in generalizing outside of the range of the examples (Bain 2000), and the domain expert needs to evaluate the limits of the learning environment. From the perspective of the domain expert, effective learning means reducing the cost of learning by improving computational performance, and reducing the cost of classification by minimizing error.

Turney (2000) presents a taxonomy of different types of costs associated with inductive learning and identifies the cost of misclassification as the most important. Misclassification is defined as assigning an unknown case to an incorrect target class. Turney (2000) identifies other cost factors that are important in evaluation, but emphasizes that all other cost can only be "rationally evaluated" in the context of classification accuracy. An important subset of these costs can be evaluated through Receiver-Operating-Characteristic (ROC) curves (Swets 1988). ROC curves plot the true

positive rate against the false positive rate of a classifier for varying thresholds, and thus

provide the feedback mechanism needed to monitor and measure improvements in the

learning environment. A ROC curve provides a tool for comparative analysis of the

performance of the learner, and is better suited for evaluation than overall accuracy alone

(Provost, *et al.* 1998).

3. SYNOPSIS

The general hypothesis of this study is that a machine learning environment using artificial neural networks can reduce the cost (negative benefit) of image classification from the perspective of the domain expert. A collaborative machine learning environment may not fully automate the process of feature classification, but even a semi-automated process may result in a significant reduction in costs. Effective and simple learning environments maximize the potential of primitive representations, and become the foundation to identify and learn higher order concepts.

Early prototypes by this author indicated that using ANNs that have a relatively small number of examples and learning cycles for supervised classification result in a predictive capacity that is comparable to manual classification (Sweet and Opitz 1999). These early trials utilized multiple experts to establish a mean and variance in expert opinion for manual classification that could be tested against the mean and variance resulting from a machine learning environment. In relationship to classification accuracy the effect of input sample size, the number of learning cycles, spatial context of inputs, and spatial proximity in selection of inputs was inconclusive in these trials. A more robust test was required to better understand and evaluate the relationship of classification accuracy to input representation and ANN model formulation.

The basic experimental framework for testing needs to begin with a design that establishes a baseline for evaluating the benefit (negative cost) of input representations or model formulations. From the perspective of the domain expert, it is only rational to use alternative input representations if the cost of misclassification is subsequently reduced (Turney 2000). Additional complexity in input representation or neural network architecture is only valid if it provides a quantifiable benefit.

The baseline classification for this study used spectral properties at the pixel level for inputs and an ANN with a single hidden layer. It learned one binary target concept. Five different binary classifications (roof tops, low vegetation, tree tops, pavement, and shoreline) were conducted over a range of sample sizes and learning cycles. There was no pre-determination of transformations on inputs. All spectral bands were present in the input representation. Both training sets and test sets were randomly selected from within a fully classified image, and it was assumed that examples were correctly classified within a tolerance known to that expert. Receiver-Operating-Characteristic (ROC) curves were used to evaluate the effects of sample size and learning cycles for each target concept.

The pixel-level baseline representations from the spectral classification experiments were evaluated against an alternative input abstraction using a foveal filter. The presumption was that a fovel representation would capture contextual information not present in pixel-

level representations. Complimenting research by Bain (2000) and Mangrich (2001), this study evaluated foveal and pixel-level representations over a wider range of target concepts, sample sizes, and learning cycles. ROC curves were used to evaluate the effectiveness of foveal representations over pixel-level representations for classification.

A third set of experiments used results from the spectral classification to evaluate concept-drift and the effect of spatial proximity on classification accuracy. Clusters of neighboring pixels provided an alternative input abstraction. Classification accuracy was evaluated as a function of distance from the input cluster. The initial spectral classification experiments guided the selection of the number of learning cycles and a sample size that would maximize the predictive capacity of clustered examples. Any bias in classification accuracy over randomly selected pixels would most likely be due to spatial proximity within and between clusters. Figure 2 provides an overview flowchart of all the experiments and analyses.

This study provides evidence of how sample size, learning cycles, and spatial representation affect feature classification as measured by classification accuracy. Both overall classification accuracy and accuracy at multiple thresholds were considered as criteria for evaluation.

Figure 2: Flowchart of investigations

Start → Acquire classified image → Determine the proportion of each class

Contextual Classification (foveal)
- Build foveal representaions of examples
- Execute contextual classifications
- Derive summay statistics and ROC curves
- Identify cases
- Evalute statistics and ROC curves
- Compare contextual and spectral classification

Spectral Classification (pixels)
- Random selection of examples
- Execute spectral classifications
- Derive summay statistics and ROC curves
- Evalute statistics and ROC curves
- Compare effect of sample size and epochs

Spectral Classification (clusters)
- Evaluate spectral classifiers — Others → End
- Best Overall
- Random selection of cluster, then select examples
- Execute spectral classifications using clusters
- Derive summay statistics and semivariograms
- Evalute statistics and semivaraiograms
- Compare effect of clustering inputs over random selection

Document results → End

14

4. METHODOLOGY

This section describes the learning environment, data set, experimental design, and evaluation methods for each of three experiments.

4.1 Learning Environment

ANNs typically have a number of learning parameters that can be set to optimize the learning environment. Recall that Foody and Arora (1997) determined that the size of the training set, the number of discriminating variables, and the size of the testing set used for evaluation were far more significant than the network architecture in determining classification accuracy. The learning parameters define the behavior of the network and the learning algorithm. The optimal network will be one that has enough hidden nodes and is not over-trained (Opitz 1997). The parameters used for the ANN learning environment in this study are straightforward, and can easily be reproduced for future experiments designed to evaluate the effect of other learning parameters on classification accuracy.

In general, learning parameters for this study were set according to experience in early exploratory work, evidence in the literature, and expert opinion. To determine the optimal value for all learning parameters would have resulted in a large number of experiments. The computational time alone makes a fully orthogonal design prohibitive.

Therefore, learning parameters were held constant except for two situations where experimental parameters dictated a change. This occurred when the number of learning cycles (epochs) was selected as one of the dependent variables in the experimental design. It also occurred when the number of hidden nodes was increased to accommodate the increase in inputs for contextual classification experiments using foveal filters.

The learning environment in this study used a stochastic backpropagation learning rule to apportion error and adjust connection weights. The experiments were conducted using the BACKPROP code and software environment (Copyright (c) 1992 by Richard Maclin, David Opitz, and Jude W. Shavlik, CS Dept., UW-Madison.) written by the machine learning group in the Computer Science Department of the University of Wisconsin-Madison. This model allows for a folded cross-validation technique that produces a confusion matrix for documenting and assessing model results. The number of cross-folds was set at 10 and percent validation was set at 0.1. Parameters defining the learning rate and momentum were set at 0.1 and 0.9 respectively (Mangrich 2001). The network units within the hidden and output layers used a standard sigmoid (logistic) transfer function to determine weights. All weights were initialized to a small random number. After learning, the resulting model is evaluated against an independent test set of examples.

4.2 Data Source

NASA's Jet Propulsion Laboratory (JPL) provided the fully classified image used for experiments in this study. JPL staff classified the image using commercial image processing techniques supplemented with manual classification and verification. Opitz, *et al.* (2000) documents the results of a visual inspection and evaluation of the classification by an independent source, and estimated a 10% error in classification across classes. The spectral data were obtained by JPL from a Positive Systems Incorporated four-band Airborne Data Acquisition and Registration (ADAR) Imager in a December 1998 over-flight of the Presidio near San Francisco, California. The ADAR Imager has sensors for blue (0.45 to 0.52 µm), green (0.52 to 0.60 µm), red (0.63 to 0.69 µm) and near-infrared (0.76 to 0.90 µm) wavelengths. Each pixel represents a ground measurement of approximately 0.5 meters. Figure 3 provides a grayscale example of a subset of the Presidio image. The image has 6500 scan lines in the Y-axis and 9200 scan lines in the X-axis, for a total 59.8 million pixels with 33,612,022 of those pixels having non-zero spectral values across all spectral bands.

Figure 3: Subset of the Presidio image in grayscale

The image was fully classified by JPL for rooftops (class 1); low-vegetation consisting of bushes and shrubs (class 2); treetops (class 3); pavement, streets and parking lots (class 4); and water or shoreline (class 5). These five classes were predominant in the image, spatially distributed, and defined the target concept or class to be learned. Since all classifications were binary, the target concept was defined as either present (1) or absent (0). For input into the learning algorithm, pixel values were normalized by dividing the spectral value for each band by the range (i.e., maximum-minimum) of that band, so that the value for each attribute was in the range [0..1] (Wilson, *et al.* 1997).

Sensors in the range of visible light typically capture a large proportion of zero values for water and shadows. Exploratory investigations indicated that a significant proportion of

18

all zero inputs severely degraded classification results. Pixels with zero values across all bands were not selected as input candidates.

4.3 Experimental Design: Spectral Classification

The spectral classification experiments used a replicated design to evaluate the performance of the learner over a range of sample sizes and epochs. The neural network consisted of 4 input units, one for each spectral band, and a hidden layer with 9 units. The output layer in the ANN had a single unit and output a real number in the range [0..1]. A threshold of 0.5 determined whether the output was a member of the target class. Each input example presented to the learner was an image pixel consisting of a target class and four normalized spectral values. For each experiment, an equal number of positive (member) and negative (non-member) examples were randomly selected from throughout the image. For consistency across experiments, negative examples were always selected in proportion to their occurrence in the full image.

As stated earlier, the literature indicates that relatively small sample sizes can produce reasonable results, but classification accuracy typically increases with an increase in the number of examples. Early experiments by this author (Sweet and Opitz 1999) indicated that a reasonable result could be obtained with a sample size as small as 250 examples. Foody and Arora (1997) used sample sizes of 10, 30, 50, and 100 pixels for training sets, and 120, 225, and 320 pixels for testing sets. In the spectral classification experiments in

19

this thesis were designed to encompass what was considered to be a small and large sample size. The experimental variable for sample size of the number of total positive and negative examples was set at 256, 1024, 4096, and 16384. The testing set size was set at 32768, which is equal to two times the maximum number of positive and negative examples across all experiments.

Early prototypes by this author indicated that an effective number of learning cycles might be in the range of 100 to 200 epochs. Since the interaction between sample size and learning cycles was unknown, the spectral classification experiments were designed to bracket this range of epochs. For each of the four sample sizes, experiments were conducted with the number of epochs set at 50, 100, 150, 200, 250, and 300. The experimental design for spectral classification resulted in 100 replicates of 5 target classes by 4 sample sizes by 6 classes of epochs for a total of 12,000 simulations.

4.4 Experimental Design: Contextual Classification

The contextual classification experiments used the same experimental design as the spectral classification experiments, but modified the neural network to accommodate a 9 x 9 foveal input representation modeled after the work of Bain (2000). The foveal representation is centered on the same target pixels selected for the spectral classification experiments. The dark lines in Figure 4 shows an example of a 9 x 9 foveal input pattern superimposed over an array of pixels. The spectral values for the eight surrounding

20

blocks were averaged for each band. The surrounding blocks were assigned the majority

target class for the nine pixels that define the block. The number of input nodes to the

ANN increased from 9 in the pixel-level representation to 68 in the foveal representation

(9 center pixels plus 8 averaged cells all with 4 spectral values). The number of hidden

units for foveal experiments was increased from 9 to 34 to accommodate the increase in

the number of input units. The target concept is represented with a value of 1. The

output is determined in the same manner as in the spectral classification experiments.
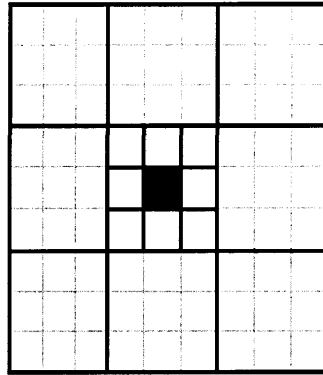
Figure 4: A 9 x 9 foveal input pattern surrounding a target pixel

Experimental parameters for sample size and epochs were identical to the spectral

classification experiments. The experimental design for contextual classification using a

foveal filter resulted in 100 replicates of 5 target classes by 4 sample sizes by 6 classes of

epochs for a total of 12,000 simulations. Table 1 summarizes the experimental

parameters for the experiments on spectral and contextual classification.

Table 1: Summary of model parameters for spectral and contextual experiments

Parameter	Spectral Classification Experiments	Contextual Classification Experiments
Momentum	0.9	0.9
Learning rate	0.1	0.1
Percent validation	0.1	0.1
Number of cross-folds	10	10

Standard net
    Input nodes
    Hidden nodes
    Output nodes
	
	4
	9
	1
		
		68
		34
		1

Input representation	single-pixel	9 x 9 foveal
Number of target concepts	5	5
Number of epochs	50, 100, 150, 200, 250, 300	50, 100, 150, 200, 250, 300
Number of examples (sample size)	256, 1024, 4096, 16384	256, 1024, 4096, 16384
Number of simulations (repetitions)	100	100
Evaluation	ROC, G-Statistic	ROC, G-Statistic
Total number of experiments	5 x 6 x 4 = 120	5 x 6 x 4 = 120
Total number of executions (experiments x repetitions)	12,000	12,000

4.5 Experimental Design: Clustering

The final set of experiments clustered the selection of inputs. A randomly selected pixel for the target class identified the center of the cluster. Single pixel samples were then selected from around the center using on a 1-dimensional nearness index called 'spatial order.' Spatial order determines the position of a pixel using a self-similar space-filling curve that iteratively segments the 2-dimensional space (Platzman, *et al.*1989). Pixels

close in 2-dimensional space will have close index values in 1-dimensional space. Figure 5 presents an illustration of one implementation of the algorithm.

	A	B	C	D

A: 0 1
B: 00 01 10 11
C: 000 001 010 011 100 101 110 111
D: 0000 0001 0100 0101 0110 0111 1000 1001 1100
1101 1110 1111
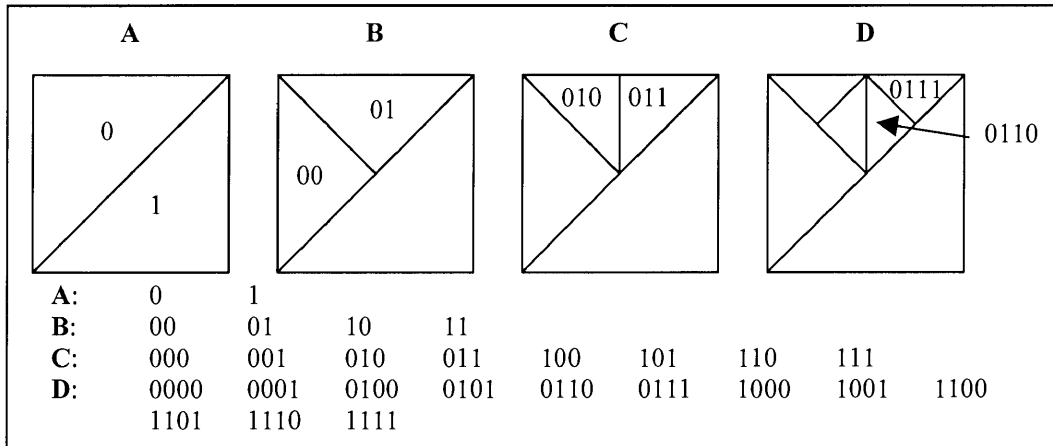
Figure 5: Example of a spatial order index algorithm

Searching for neighboring pixels was computationally efficient when the 1-dimensional spatial order index was used as a distance measure. Examples were selected from around the center pixel using an incremental search window. The search window increased in dimension until the target sample size for the number of positive examples was satisfied. Negative examples were selected in equal proportion to the number of positive examples, but the proportion of negatives from each class differed from the pixel-level spectral classification. Selecting negative examples in proximity to the center of the cluster could not result in a random selection for each class in proportion to their occurrence within the full image, as was required for the spectral classification experiments. The center pixel and the surrounding collection of positive and negative pixels defined the sample cluster.
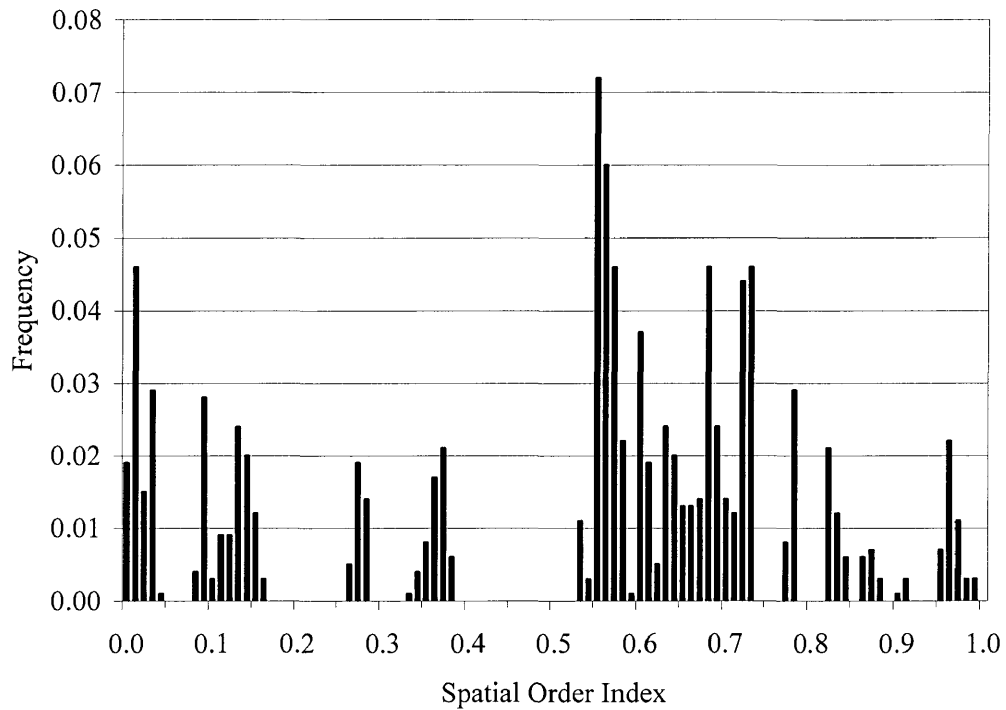
Figure 6: Frequency distribution of spatial order index for rooftops

Figure 6 displays a graph of the frequency distribution of rooftops by spatial order index for all rooftops within the image. Other classes displayed a similar distribution pattern. Figure 7 illustrates a sample cluster with all potential positive examples appearing in dark gray, and selected positive examples appearing in light gray. Further discussion on parameters for the cluster experiments are presented in the results, since these parameters are based on results from the spectral classifications.
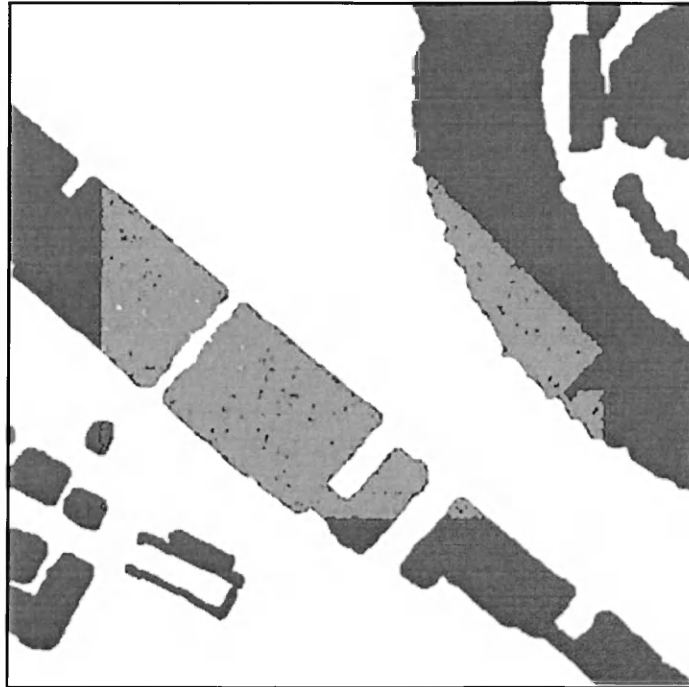
Figure 7: Example of a cluster of positive examples for rooftops

4.6 Evaluation Methods

The objective in applying ANNs to image classification is to have a high probability of

correctly detecting the target class, and different formulations of the learning

environment will produce different outcomes. To select among alternate formulations,

the domain expert must have tools for evaluating the performance of the learner. If

differences in outcomes are to be meaningful, they must be measurable and identify the

learning conditions that minimize variability in results across repeated simulations. In

the case of ANNs, a reduction in the degree of variability means identifying a result that

generalizes well. It is possible to have two resulting ANN models with the same

predictive power, but the model with less variability over a range of learning conditions is preferred.

Effective evaluation tools for a comparative analysis include Receiver-Operating-Characteristic (ROC) curves, statistical comparisons of ROC curves, and statistics that compare overall mean accuracy and standard error (Provost, *et al.* 1998). ROC curves plot the false positive rate against the true positive rate. Both rates result in numbers ranging from 0 to 1. The true positive rate is the ratio of correctly classified pixels to the total possible pixels for the target class. The false positive rate is the ratio of pixels falsely classified as the target class to the total of all non-target pixels. At the origin of the ROC graph the classifier finds no positives, and at point (1,1) everything is positive. Classifiers that fall along the diagonal line from (0,0) to (1,1) are random guessers. The best classifiers operate at point (0,1), which indicates a high rate of true positives and low rate of false positives. The ROC curve can be used to select a particular threshold with a known tradeoff between the true and false positive rates. ROC curves provide feedback to the domain expert on the relative contribution of model components.

An ROC curve represents a distribution, and a comparative analysis is typically accomplished by evaluating the area under the curve. A more robust comparison of ROC curves can be accomplished using the *G*-Statistic (Sokal and Rohlf 1981). The *G*-Statistic is a replicated goodness-of-fit test for a comparative analysis of two distributions. It determines if the ROC curves are significantly different. Where ROC

curves are significantly different, the *G*-Statistic provides a means to use paired comparisons to determine at what threshold portions of the curves are significantly different. According to Foody and Arora (1997) a "good-fit" model is regarded as the evaluation of the effects of factors at a desired level of significance, generally $\alpha = 0.05$ percent.

ROC curves can be extremely valuable in interpreting the relative contribution of alternative approaches. The significance of the difference and the magnitude of the effect are important when evaluating multiple ROC curves. ROC curves can be a basis for quantifying and interpreting some of the types of costs, and in quantify the difference in costs between classifiers (Drummond and Holte 2000, Turney 2000). Turney (2000) emphasized that cost and benefit (negative cost) can only be valued in the context of the classification objective as defined by the domain expert. Some examples of costs to consider in evaluating the performance of classifiers are:

- Conditional error costs: The costs of a certain error may be conditional on the circumstance, such as the selection of thresholds. ROC curves provide an evaluation tool with measures for this case.

- Cost of test: The domain expert can only rationally determine whether it is worthwhile to pay the costs of a test when the cost of the misclassification errors is known. This requires that improvements be measurable, and interpreted in the context of the problem domain.

- Cost of teacher: Given an unlimited supply of unclassified examples, as is the case with imagery, it is expensive to determine the correct class of an example. The cost of teacher increases with increased complexity due to an increased partitioning of the problem space for both the teacher and learner.

- Cost of computation: The complexity of size, structure, time, memory space,

training, and testing should all be minimized.

- Cost of cases: The number of cases required to develop a given model with a known rate of misclassification reduces cost should be minimized.

- Human-Computer Interaction Cost: The number of parameters and the cost of incorporating domain knowledge should be minimized.

- Cost of instability: The learning algorithm must be stable and repeatable, and must emphasize the tradeoff between generalization and specialization.

It is important to emphasize that costs are relative to the domain, and a learning environment that provides the domain expert with tools to evaluate costs can offer significant advantages.

A semivariogram and a scatterplot of error by distance were used to evaluate the effect of clusters of input pixels on classification accuracy. Both methods illustrate the effect of spatial proximity on classification accuracy. A semivariogram provided the means to evaluate the relationship between proximity and classification accuracy on a pixel-to-pixel basis. A scatterplot provided an evaluation of overall classification accuracy for clusters in relationship to proximity of the cluster used as a training set.

A semivariogram is a function that describes the spatial dependence of error. It identifies the distance at which there will be no spatial dependence on a point (pixel) and its neighbor. Curran (1988) offers an in-depth discussion on the application of semivariograms to remote sensing. Curran suggests that high-resolution imagery is

preferred since spatial dependencies often occur within a small distance from the source pixel. The semivariograms developed for this study used Euclidean distance as the distance measure for determining lag. Lag is the distance between sampling pairs of pixels. In a semivariogram the lag is plotted against the semivariance. Semivariance is the average variance of the differences between all sample pairs.

5. RESULTS

5.1 Results from Spectral Classification Experiments

The spectral classification experiments were based on 100 training sets for each

combination of model variables and values listed below:

Number of epochs: 50, 100, 150, 200, 250, 300
Sample size (pixels): 256, 1024, 4096, 16384
Target classes: rooftops (1), low vegetation (2), treetops (3), pavement (4), and shoreline (5)

An overall mean classification accuracy and standard error was calculated using a

standard threshold of 0.5. Output values great than or equal to 0.5 were assigned to the

target class of the model. To support the evaluation of effect over a range of thresholds

using ROC curves, output values were summarized using threshold increments of 0.1 for

the range [0..1].

A simple exploratory test was conducted to learn more about the effect of the number of

epochs outside the upper bound of 300 epochs chosen for the spectral classification. The

test evaluated a range of epochs in increments of 100 for the range [100..1000] using a

sample size of 4096 for all classes. Figure 8 illustrates results for the rooftop class.

Other classes demonstrated a similar pattern. Each point on the graph represents overall

mean classification accuracy for 100 replicates for a given epoch. Vertical bars indicate

plus and minus one standard error. A pair-wise comparison of means was conducted

using the T-Method (Sokal and Rohlf 1981). This compares the difference of means

against a minimum significant difference (MSD). The MSD for a pair-wise comparison

is a critical value of 4.063 ($\alpha = 0.05$) multiplied by average standard error. Differences

greater than the MSD value are significant. The exploratory test indicated that epochs of

300 and greater were not significantly different ($\alpha = 0.05$) except for epoch 900, which

performed more like epochs 100 and 200. The pattern of mean accuracy and the increase

in standard error at epochs greater than 300 could be an indication of over-fitting the

ANN model to the examples. The exploratory test supports the selection of 300 epochs

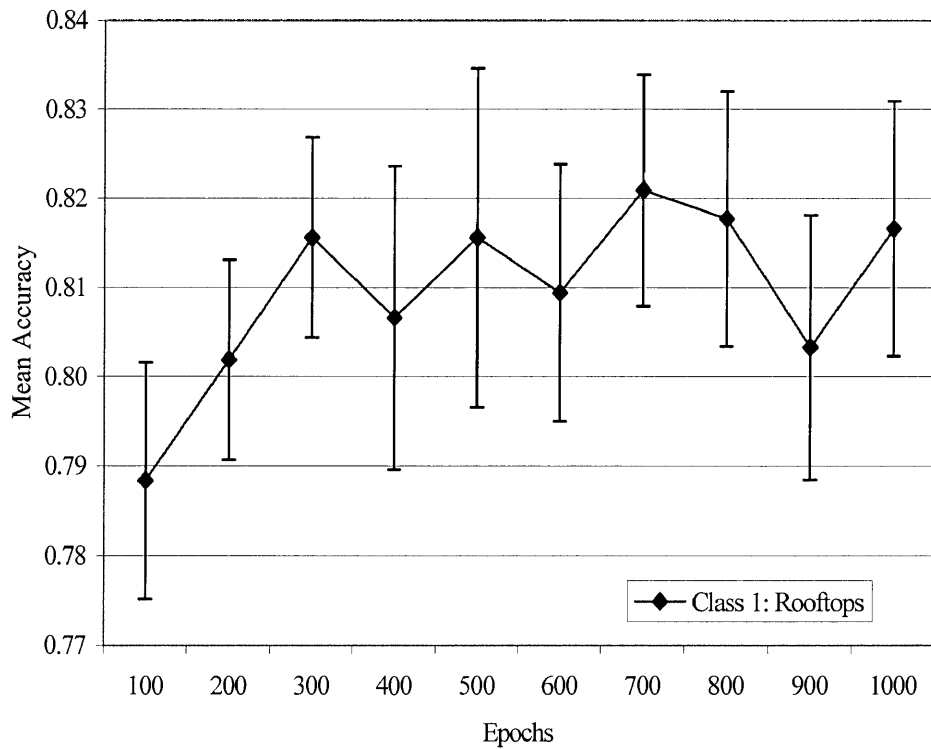as a reasonable parameter setting for the spectral classification experiments.

Figure 8: Mean classification accuracy and standard error by epochs for rooftops

In general, as sample size and number of epochs increased the spectral classification experiments resulted in higher mean accuracies and lower mean standard errors. Table 2 lists the minimum and maximum mean classification accuracy and respective standard error across all sample sizes and epochs for each class. Table 2 also lists the average mean and average standard error for each target class across all simulations. A mean of 1.0 would indicate there are no errors in classifying the target concept.

Table 2: Mean accuracy and standard error by target class for spectral classification

Target Class	Min. Mean (S.E)	Avg. Mean (Avg. S.E.)	Max. Mean (S.E.)
1. Rooftops	0.6574 (0.0429)	0.7770 (0.0191)	0.8267 (0.0095)
2. Low vegetation	0.6605 (0.0419)	0.7539 (0.0260)	0.8323 (0.0156)
3. Treetops	0.6360 (0.0442)	0.7128 (0.0239)	0.8007 (0.0114)
4. Pavement	0.6517 (0.0369)	0.7610 (0.0219)	0.8067 (0.0162)
5. Shoreline	0.7087 (0.0374)	0.8828 (0.0240)	0.9655 (0.0162)

A pair-wise comparison of means, using the T-Method described above, identified breakpoints for each target class at which sample size and number epochs are no longer significantly different. Table 3 lists the breakpoint values. A sample size or epoch larger than the breakpoints listed in Table 3 resulted in a classification that was at least as good, but the difference was not significant. A given class may have more than one sample size and epoch listed in the table. This is because there was an interaction between sample size and epochs, and a larger sample size with fewer epochs may have performed as well as a smaller sample size with more epochs. An analysis of standard error of the mean indicated that a higher number of epochs resulted in a lower and more uniform standard

error, which is a preferred outcome for a generalized model. Figure 9 displays an example of trends in standard error by epochs and sample size for rooftops.

Table 3: Significant sample size and epoch breakpoints for each target class

Target Class	Sample size	Number of Epochs
1. Rooftops	4096	50
2. Low vegetation	4096	150
2. Low vegetation	16384	50
3. Treetops	4096	250
3. Treetops	16384	50
4. Pavement	1024	200
4. Pavement	4096	50
5. Shoreline	1024	50
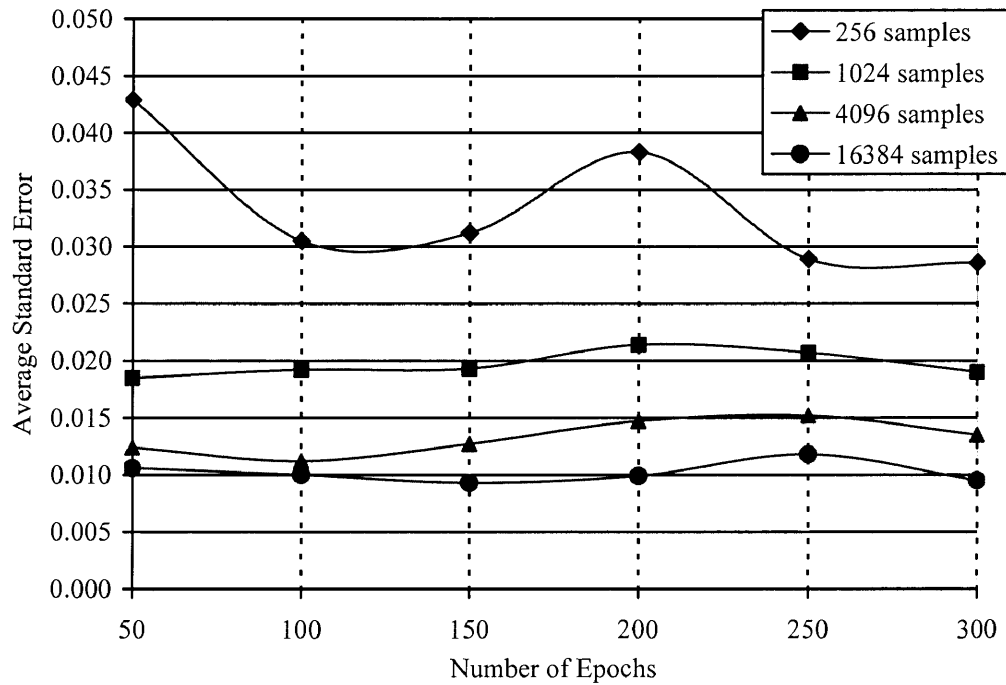
Figure 9: Standard error by epochs by sample size for rooftops

An analysis of overall classification accuracy provided information on the interaction of sample size and epochs for each target class using a threshold of 0.5. To evaluate the effect of sample size and epochs over a range of thresholds, ROC curves were compiled for all combinations of model variables represented in the experimental design. The following figures illustrate examples of the value of ROC curves for interpreting results.
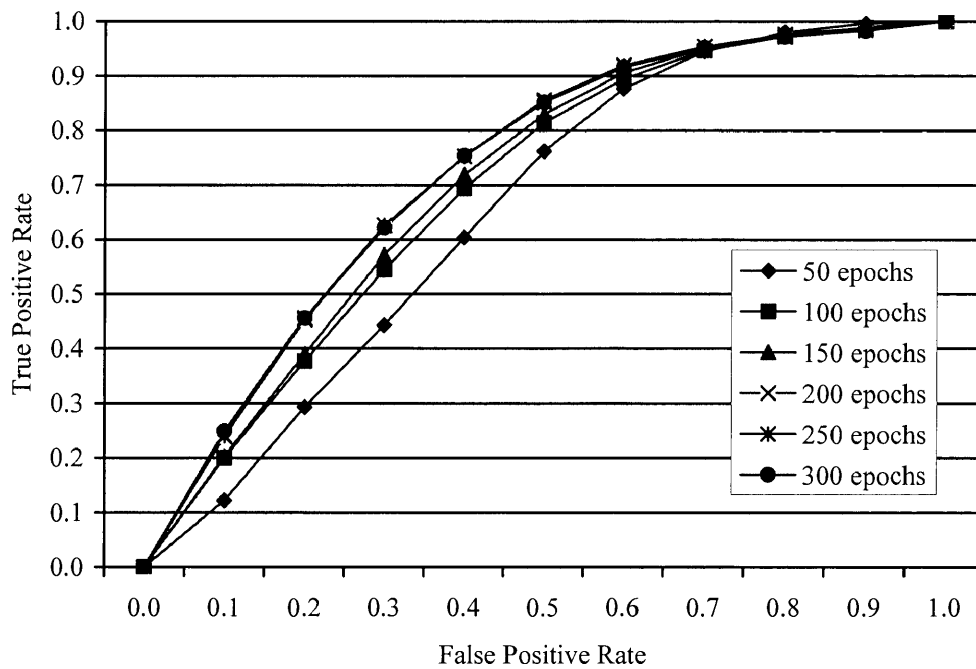
Figure 10: ROC curves for sample size of 256 across all epochs for rooftops

Figure 10 shows an ROC curve for each epoch for a sample size of 256 examples for the target class of rooftops. Deriving a G-Statistic ($\alpha = 0.05$) at each threshold provides a pair-wise comparison of statistical significance. The G-Statistics for the ROC curves in Figure 10 state that there is no significant difference ($\chi^2_{.05[1]} = 3.841$) in classification

accuracy across epochs given a range of thresholds for a sample size of 256 for rooftops

(Class 1). This result supports the breakpoint of 50 epochs for rooftops in Table 3. A

comparison of overall mean accuracy at a threshold of 0.5 confirms that there is no

significant difference between epochs.

Figure 11 shows an ROC curve for each sample size for an epoch of 150 for rooftops

(Class 1). The *G*-Statistics for the ROC curves in Figure 11 state that there is no

significant difference between sample sizes of 4096 and 16384, and a sample size of

1024 performed as well as of the other sample sizes. A comparison of overall mean

accuracy at a threshold of 0.5 confirms these results.

Figure 11: ROC curves for epoch 150 across all sample sizes for rooftops

Access to a fully classified image and multiple target classifications provided for

comparisons that are not typically available. To further evaluate the interaction between

classes, each resulting model for each class was employed as a predictor for each of the
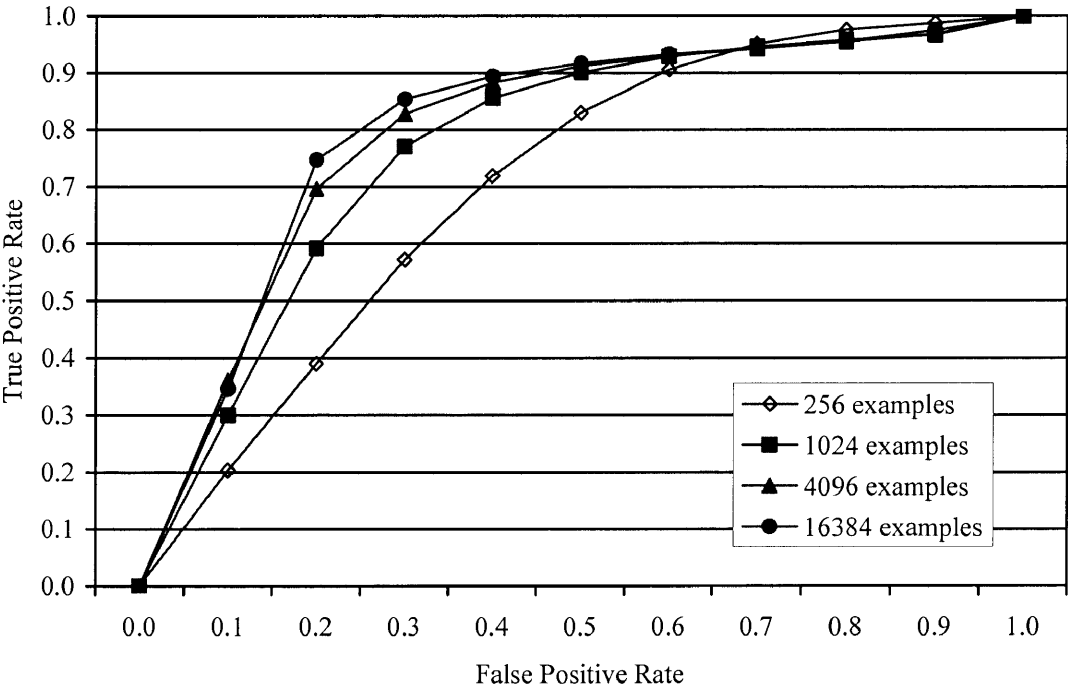
other classes. That is, the resulting model for rooftops (Class 1) was used as a predictor

for treetops, low vegetation, pavement, and shoreline (Classes 2 through 5). These are

the negative cases. If there were true spectral separation between classes at the pixel-

level, this analysis would show if any given classifier was very poor at classifying any of

the other class. Further analysis of the spectral classification experiments revealed some

interesting results:

- There was some confusion between all target classes at a sample size of 256
  across all epochs. A sample size of 256 examples is insufficient.

- In cases where there was some separation, class separability improved with an
  increased sample size or increased number of epochs.

- In general, a smaller sample size and fewer epochs were required to detect class
  separability than to detect a difference in overall mean accuracy.

- Rooftops (Class 1) and pavement (Class 4) were not separable across the full
  range of sample sizes and epochs. In general, the models for rooftops were good
  predictors of pavement, and the pavement models were good predictors of
  rooftops.

- Low vegetation (Class 2) and treetops (Class 3) were not separable across the full
  range of sample sizes and epochs. In general, the models for low vegetation were
  good predictors of treetops, and the models for treetops were good predictors of
  low vegetation.

The results are significant because they tell the domain expert that the input

representation for the spectral classification should be altered, since increasing the sample

size and number of epochs had no significant effect on resolving the spectral confusion between rooftops and pavement, and between low vegetation and treetops.

Revisiting the breakpoints in Table 3, there is some evidence of the impact of this spectral confusion. Shorelines (Class 5) occupy a fairly narrow spectral niche and the spectral distribution differs significantly from the other classes. The breakpoint for sample size and epochs for the shoreline class is low when compared to other classes. Rooftops and pavement have higher breakpoints, and low vegetation and treetops have the highest breakpoints. The higher breakpoints in these latter cases may be indicative of the degree of spectral confusion, and the difficulty the learner had in arriving at a plausible hypothesis.

ROC curves were valuable in identifying both the presence and degree of spectral confusion. As an example, Figure 12 shows an ROC curve for rooftops as a predictor of all other classes. This type of ROC curve reflects negative cases to the positive side, as if predicting for the negative case. The classifier's output is negatively correlated with the target class (Langdon 2001). This type of ROC curve is labeled as an r-ROC curve to differentiate it from a standard ROC curve. As an r-ROC curve approaches the upper-left it demonstrates better class separability. This interpretation of "goodness" is consistent with a standard ROC curve. An r-ROC curve that approaches the (0,0) to (1,1) diagonal demonstrates poor class separability, just as a standard ROC would depict this diagonal

line as indicative of a poor classifier. In Figure 12 the line for pavement (Class 4) illustrates the confusion with rooftops (Class 1) while retaining good separability with low vegetation, treetops, and shoreline (Classes 2, 3, 5). In Figure 13 the line for treetops (Class 3) shows confusion with low vegetation (Class 2), although not as pronounced as in the case shown in Figure 12.
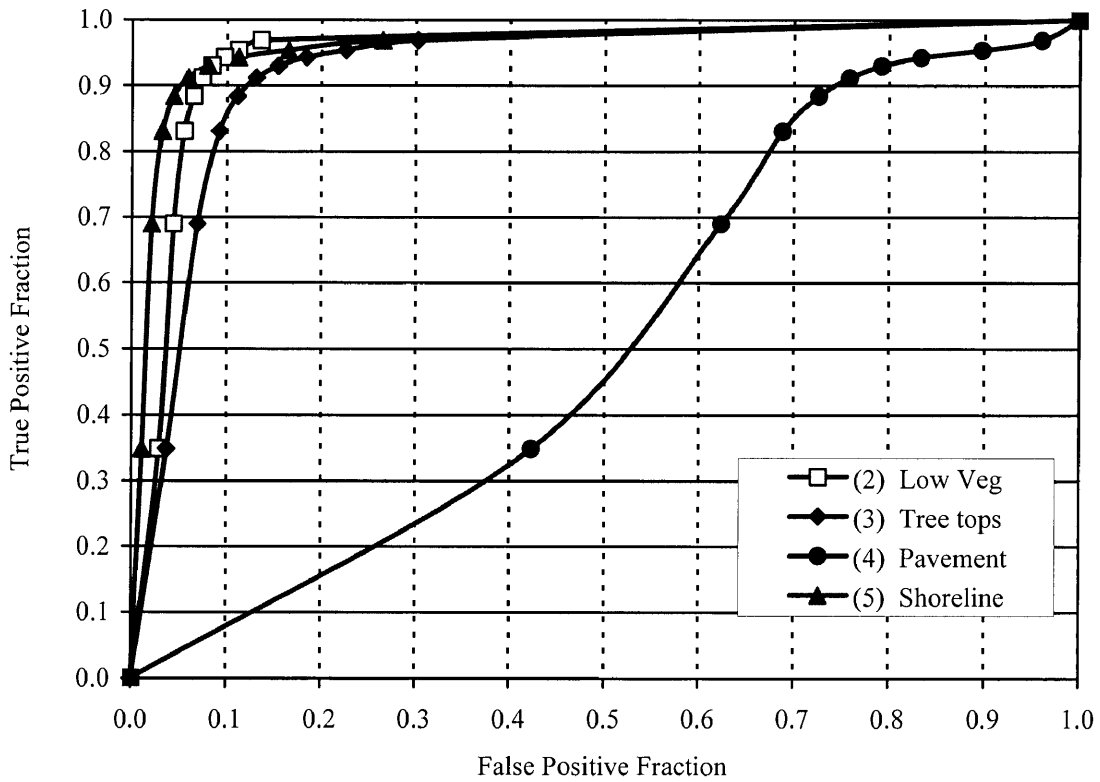
Figure 12: Confusion between rooftops and pavement
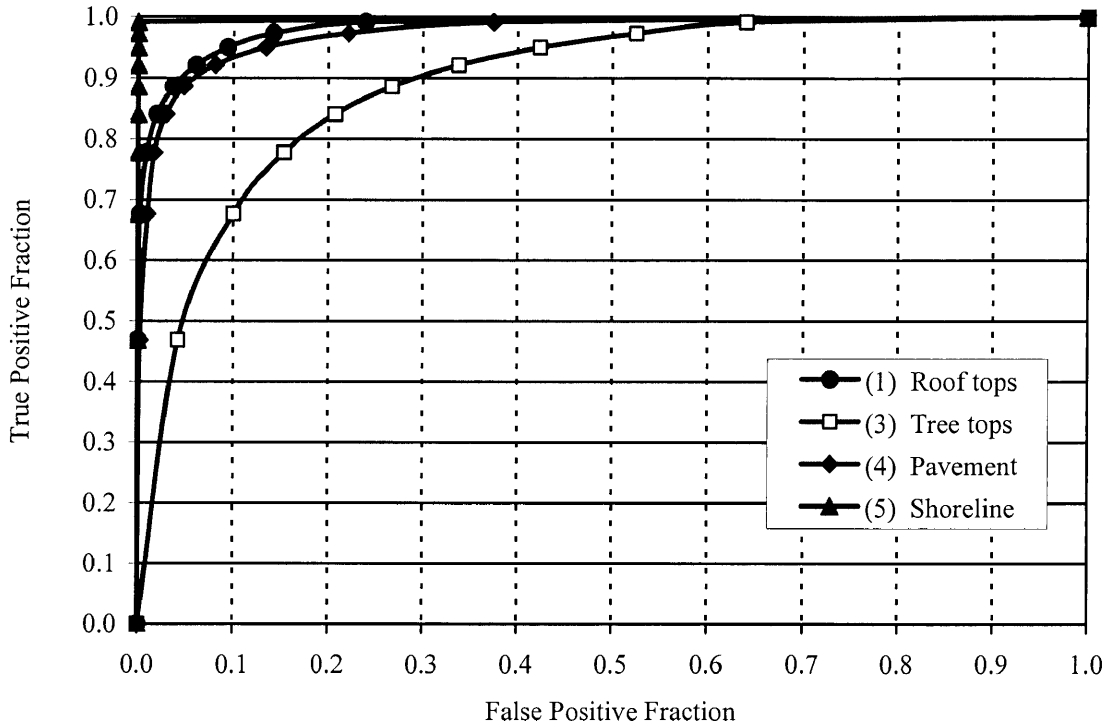
Figure 13: Confusion between low vegetation and treetops

A retrospective analysis of the spectral signatures also shows confusion. Figure 14 and

Figure 15 graph a frequency distribution of the number of pixels by Euclidean distance

for the cases identified above. Euclidean distance in spectral space was computed as the

square root of the sum-of-squares of the four spectral values for each pixel for each class.

This is equivalent to a Euclidean distance calculated as an offset from the origin. This

distance calculation collapses the spectral values of the four spectral bands into a single

number for each pixel. Figure 14 and Figure 15 graph a frequency distribution of the

distance values and display how two different target concepts share "spectral space." A

G-Statistic calculation indicated that these curves were significantly different ($\alpha = 0.05$).

In their classification process the Jet Propulsion Lab noted the spectral confusion between

rooftops (Class 1) and pavement (Class 4), and between low vegetation (Class 2) and

treetops (Class 3) (Mangrich 2001, Opitz, *et al.* 2000). The ANN learner did not

overcome that confusion as evidenced by Figure 12 and Figure 13. This implies that an

additional input or a transformation on the inputs may be required prior to submitting

examples to the ANN learner.
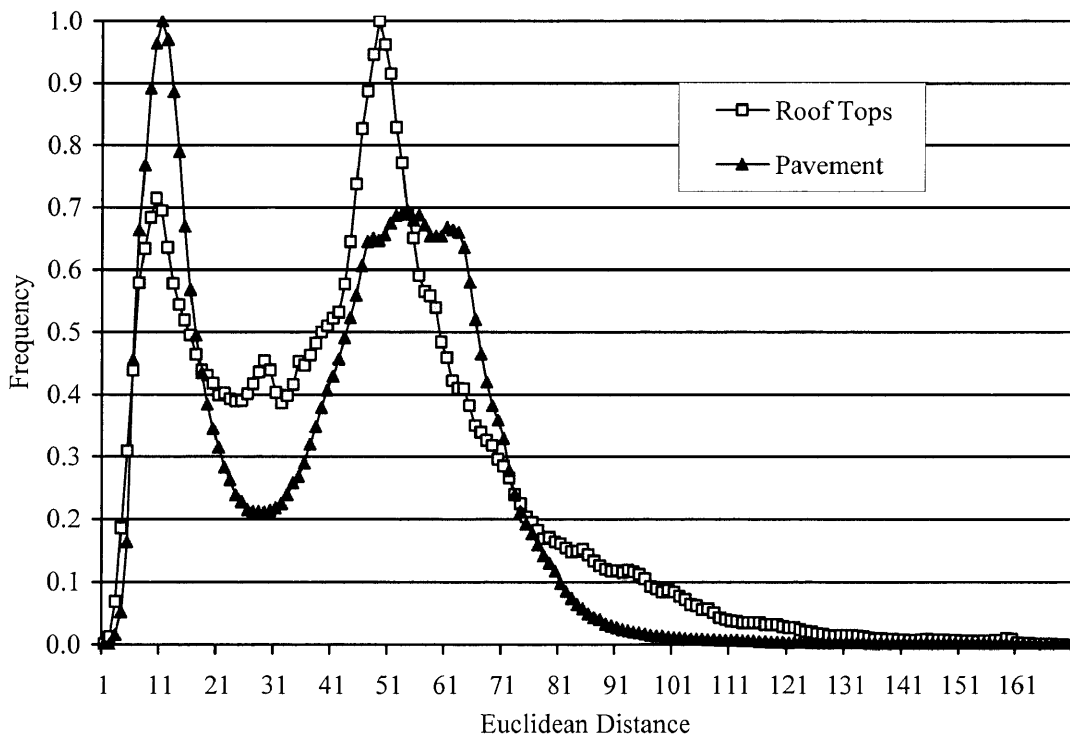
Figure 14: Frequency of spectral values for rooftops and pavement
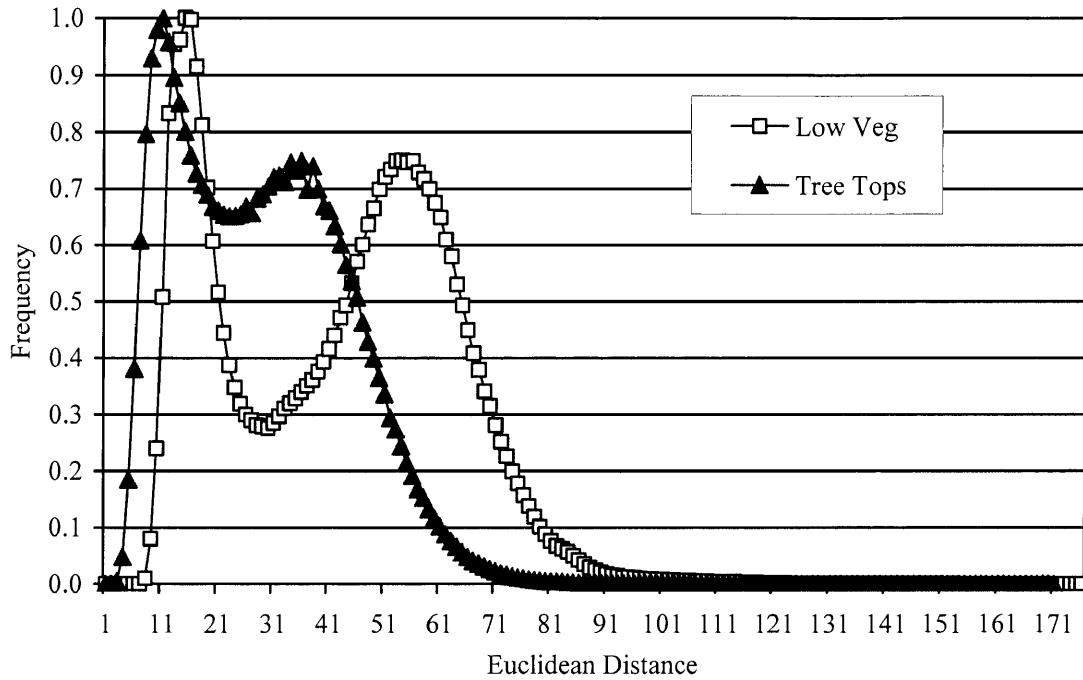
Figure 15: Frequency of spectral values for treetops and low vegetation

5.2 Results from Contextual Classification Experiments

To support a comparative evaluation the spectral and contextual classification experiments used the same experimental design. Evaluation of the contextual classifications focused on those cases in the spectral classification experiments where there was evidence of poor separability between classes. If a contextual classification using a foveal filter did not significantly improve overall accuracy or class separability, then the expert would not be inclined to incur the added cost of using foveal representations. This is the cost of tests decision described by Turney (2000). An analysis of the contextual classification experiments showed some interesting results:

41

- In the spectral classification experiments there was confusion between all classes at a sample size of 256. In general, the foveal representation resulted in a better classification at smaller sample sizes, but significantly better only with epochs of 150 or greater.

- The foveal representation significantly increased standard error in the output over pixel-level representations.

- With foveal representations, there was greater separation between rooftops and pavement when rooftops were the target class. This was not true when pavement was the target class.

- With foveal representations, there was greater separation between treetops and low vegetation when treetops were the target class. This was not true when low vegetation was the target class.

The following example compares some results from the contextual classification with results from the spectral classification for cases where class separability was poor (rooftops and pavement, and low vegetation and treetops). Table 4 lists the overall mean accuracy and standard error for the spectral and contextual classifications for a sample size of 4096 and an epoch of 150.

Table 4: Comparison of mean and standard error for pixel and foveal representations

Target Class	Pixel Mean (S.E.)	Foveal Mean (S.E.)
1. Rooftops	0.8074 (0.0127)	0.7854 (0.1890)
2. Low vegetation	0.7928 (0.0234)	0.6726 (0.1916)
3. Treetops	0.7831 (0.0112)	0.8235 (0.2216)
4. Pavement	0.8001 (0.0161)	0.7717 (0.1884)
5. Shoreline	0.9459 (0.0207)	0.9753 (0.1205)

In general, the overall mean accuracy decreased with the foveal representation. More importantly, standard error increased significantly. Treetops and shoreline represent two cases where the overall mean accuracy improved with the foveal representation. In the

case of treetops, the increase in standard error masks any difference in the means. The

shoreline results may have improved simply because of an increase in the number of

redundant inputs. The shoreline class has a relatively pure spectral signature compared to

the other classes, and is spatially distinct from the other classes.

When rooftops (Class 1) were used as a predictor for the negative case of pavement

(Class 4), the overall mean classification accuracy for rooftops dropped from 0.7213 to

0.6308. That is an indication that foveal representation provided greater class

separability between rooftops and pavement. This is illustrated in the r-ROC graph in
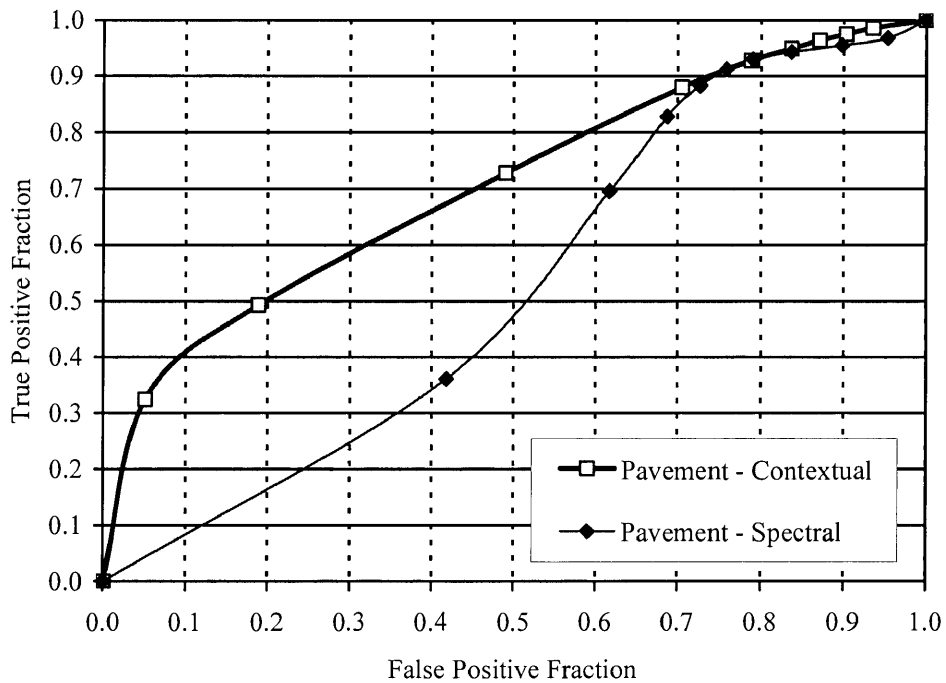
Figure 16.

Figure 16: r-ROC for contextual versus spectral, rooftops and pavement

In interpreting the r-ROC in Figure 16 recall that a line approaching the [(0,0),(1,1)]

diagonal is a good predictor of the comparative class. In this case, using a pixel-level

representation, the model for rooftops (Class 1) is a good classifier of pavement (Class 4).

As a curve approaches the upper-left of the graph, it indicates greater class separability.

In the case presented in Figure 16 the foveal representation did improve class

separability, but at the cost of a slight reduction in overall accuracy and a significant

increase in standard error. The foveal representation did not generalize as well as the

pixel representation, but was better at minimizing confusion between some classes.

Figure 17, Figure 18, and Figure 19 illustrate r-ROC curves for the three other cases

where there was an indication of poor class separability. This occurred with low

vegetation and treetops, and its inverse, and with the inverse of the previous case of

rooftops and pavements. Figure 17 shows the foveal representation did not improve class

separability over the pixel-level representation when classifying for pavement. Figure 18

and Figure 19 show an inverse relationship to each other. In Figure 18 the pixel-level

representation performed better at class separation when classifying for low vegetation

(Class 2). In Figure 19 the foveal representation slightly improved class separability

when classifying for treetops (Class 3). At epochs of 50 and 100, there was little or no

gain in class separability. The outcomes described above only became apparent at epochs

of 150, and outcomes for epochs greater than 150 were not significantly different from

those at 150. Sample size followed a similar pattern. Sample sizes smaller than 4096

examples had poorer results, and the rate of degradation increased as sample size
decreased. Sample sizes larger than 4096 examples had better results, but the rate of
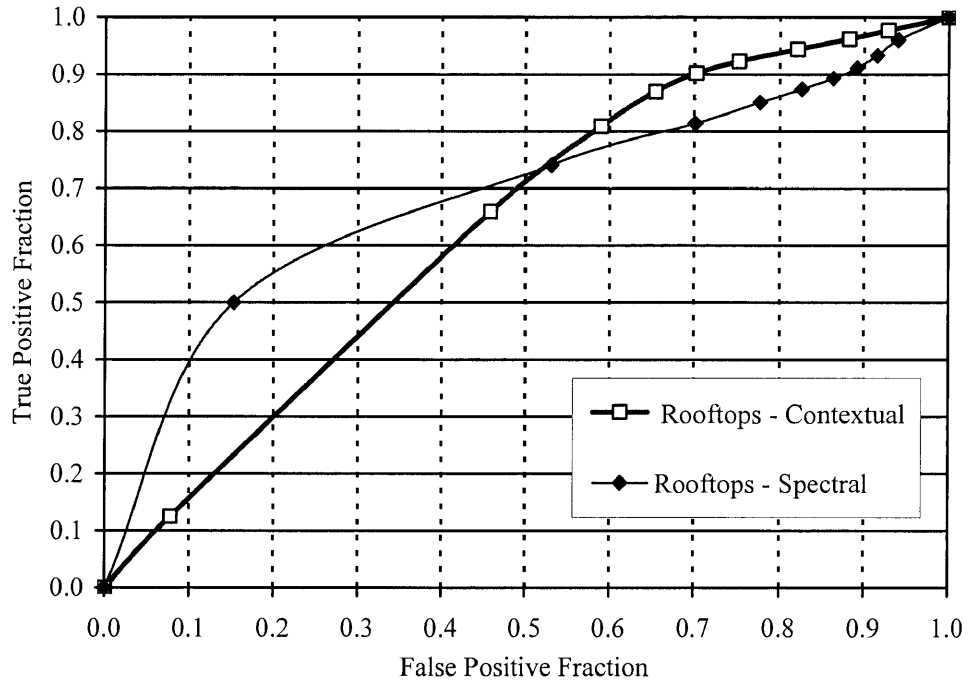improvement decreased as sample size increased.

Figure 17: r-ROC for contextual versus spectral, pavement and rooftops
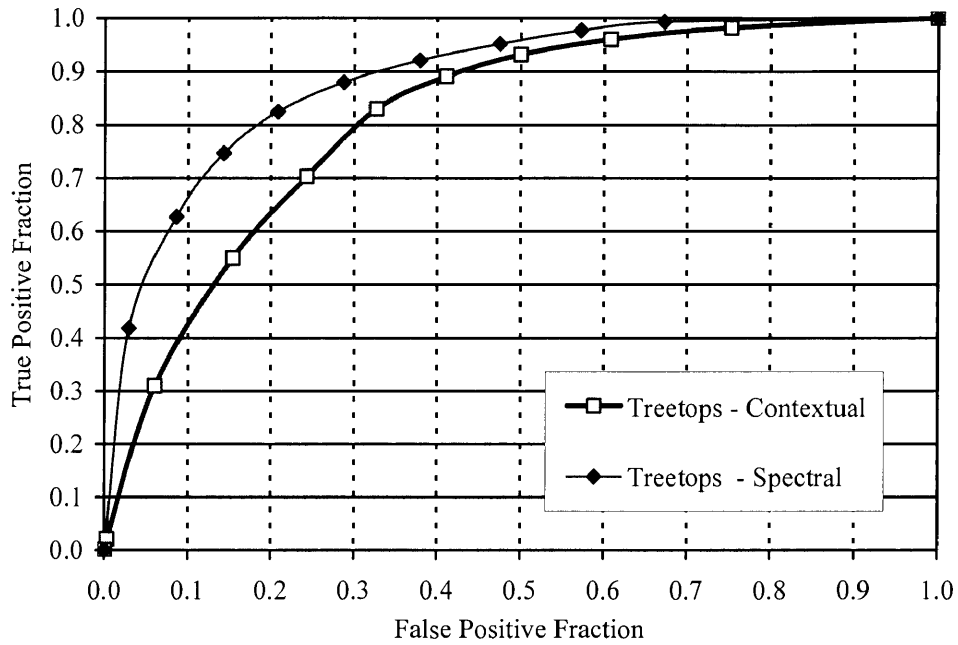
Figure 18: r-ROC for contextual versus spectral, low vegetation and treetops
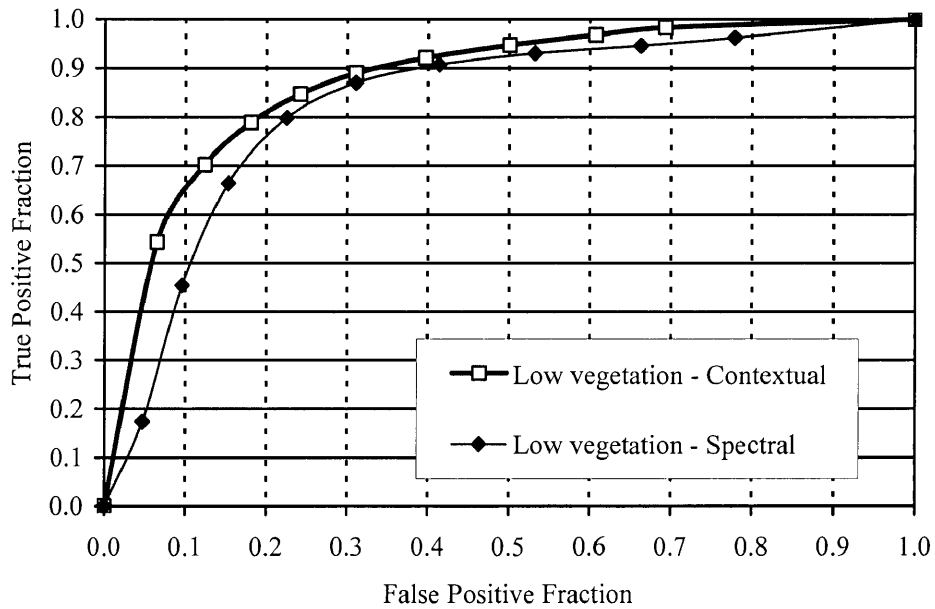
Figure 19: r-ROC for contextual versus spectral, treetops and low vegetation

In general, from the viewpoint of the domain expert, foveal representations add to the cost of classification due to the added cost of computation, instability, and misclassification. The graphs in Figure 16 through Figure 19 represent the mean of 100 independent models. Recall that the standard error in mean accuracy is significantly higher when using foveal rather than pixel-level representations as inputs. This means that the foveal representations had a broader range of outcomes. Some contextual classification models were much better and some much worse than the spectral classification models. Additional analysis would be required to determine the underlying conditions that resulted in very good and very poor contextual classifiers.

5.3 Results from Cluster Experiments

While the foveal experiments attempted to capture information on spatial context in support of classification, the final set of experiments focused on the effect of spatial proximity. The spectral classification experiments indicated that a reasonable result could be obtained using individual pixels randomly selected from within the image. A more efficient selection strategy is to have a domain expert identify a cluster of pixels that represent the target concept. The clustering of both positive and negative examples resulted in a selection of examples that is not as random as the approach used for spectral classification experiments.

The spectral classification experiments indicated that a sample size of 16384 and 300

epochs was overall a good-fit model across all target classes. Using these experimental

parameters for the experiments with clusters minimize differences in classifiers. Any

difference in classification accuracy should be due primarily to the clustering of pixels or

the proximity of target clusters to training clusters. The spectral classification

experiments indicated that 20 replicates were sufficient to provide a reasonable estimate

of the overall mean and standard error for classification accuracy for a sample size of

16384 examples and 300 epochs. For each of the 5 target classes a total of 20 clusters

were identified. Table 5 summarizes the experimental parameters for the cluster

experiments.

Table 5: Summary of model parameters for cluster experiments

Parameter	Cluster experiment
Momentum	0.9
Learning rate	0.1
Percent validation	0.1
Number of cross-folds	10

Standard net
    Input nodes
    Hidden nodes
    Output nodes
	
	4
	9
	1

Input representation	Cluster
Number of target concepts	5
Number of epochs	300
Number of examples (sample size)	16384
Number of simulations (repetitions)	20
Evaluation	semivariogram, scatterplot
Total number of experiments	5 x 1 x 1 = 5
Total number of executions (experiments x repetitions)	100

For the target class each cluster of examples was used as a training set, and one model was generated for each cluster. This resulted in a total of 20 models. The resulting model from each cluster was applied to the other 19 clusters. This provided 19 test sets for each model. Concentrating all examples in one training cluster is the most extreme case for clustering the selection of inputs. An overall accuracy for each target class was derived from aggregating the results from each of 20 different models and their associated test sets. Table 6 summarizes training set and test set overall mean accuracy and standard error for each class. Also, the test set summary from the spectral classifications in Table 2 is reiterated in Table 6. The cluster representation has a lower overall mean accuracy and higher standard error than the spectral classification experiments. It may be of interest to note that the cluster training set had a higher standard error than the pixel test set. While the number of replicates is indeed smaller than the pixel-level experiments, the increase in standard error may be an indication of a high level of within cluster variability due to the change in input selection.

Table 6: Mean and standard error for training and test set clusters

Target Class	Training Set – Cluster		Test Set – Spectral		Test Set – Cluster	
	Mean	S.E.	Mean	S.E.	Mean	S.E.
1. Rooftops	0.8742	0.0669	0.8267	0.0095	0.7036	0.1095
2. Low vegetation	0.9438	0.0328	0.8323	0.0156	0.7134	0.1495
3. Treetops	0.8740	0.0636	0.8007	0.0114	0.7176	0.0976
4. Pavement	0.9034	0.0607	0.8052	0.0111	0.7552	0.1280
5. Shoreline	0.9948	0.0132	0.9645	0.0163	0.9232	0.1565
	n = 20		n = 100		n = 380	

Figure 20 is a graph of training set error along a single transect within a cluster when predicting for the target class of rooftops. The origin of the graph is the center of the cluster, and the right-most value on the x-axis is the edge of the cluster. This graph indicates that the model was better at predicting a certain subset of the rooftop class, and did not generalize well to all types of rooftops with the class. Evaluating both training set and test set classification accuracy within a cluster may provide some insight as to how to further partition the problem set for learning.
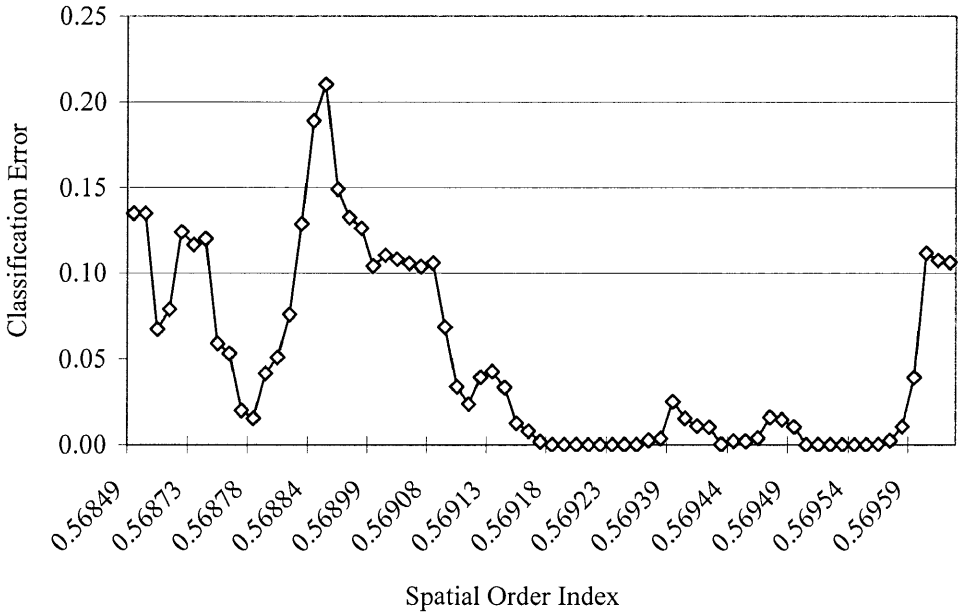

Figure 20: Classification error for rooftops by spatial order index within a cluster

Semivariograms were developed for the cluster test sets to illustrate the effect of spatial proximity. The lag value in the semivariograms is Euclidean distance between pair-wise selections of pixels. A pixel pair was composed of a training set pixel within a cluster

50

and a randomly selected test set pixel from within the image. Semivariance is the average variance of the difference between pairs, and there were over eight thousand pairs per cluster. To illustrate the results, the difference in variance was averaged within 20 different bins of Euclidean distance. Figure 21 and Figure 22 illustrate the results with one line on the graph for each of the 20 clusters. The graphs indicate a very weak and varied relationship between classification error and lag. Curran (1988) would classify these as aspatial semivariograms. The resulting models did generalize well to pixels that were not within close proximity to the cluster, but there was a high degree of variability between models.
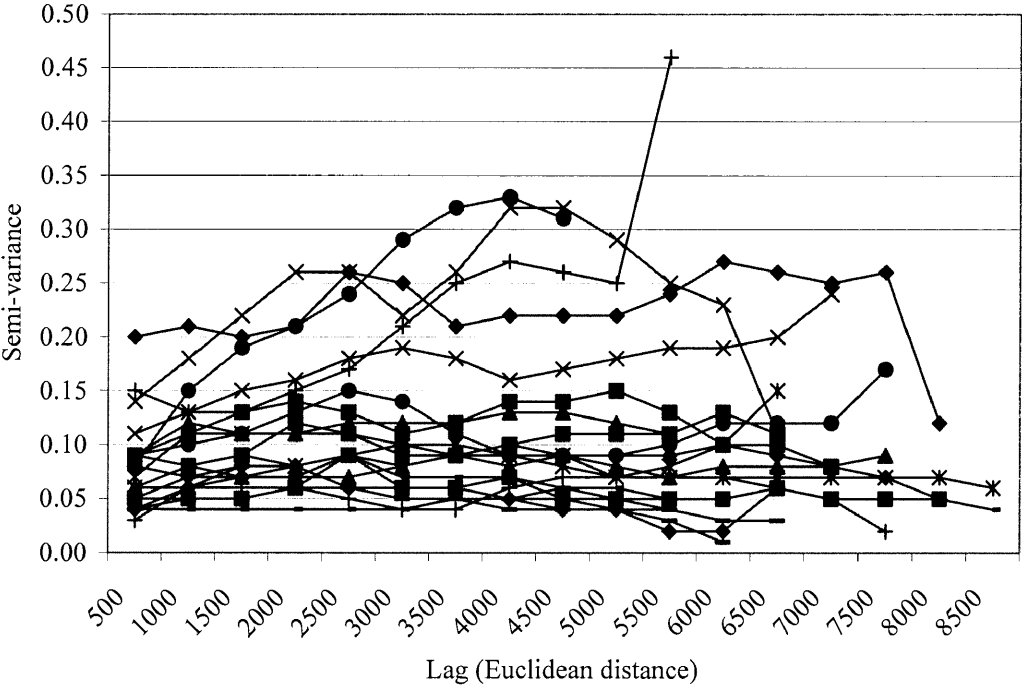
Figure 21: Semivariogram for rooftops using clusters
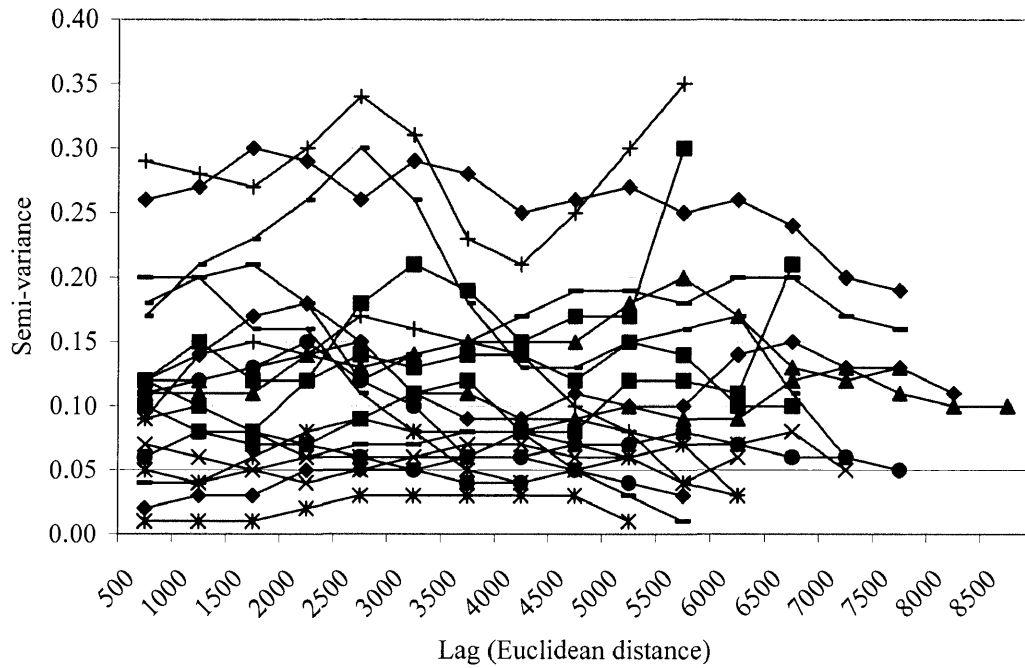
Figure 22: Semivariogram for pavement using clusters

These results from the clustering experiments indicated that selecting inputs in close

proximity to each other reduced overall classification accuracy and increased standard

error. Figure 23 plots the difference in the overall mean accuracy between a training

cluster and test set clusters as a function of the distance between the centers of clusters.

Spatial order index was used as the distance measure. Figure 23 also highlights the values

for just the second cluster of twenty total clusters. The scatterplot illustrates that there is

no apparent bias in prediction of the target class due to spatial proximity.
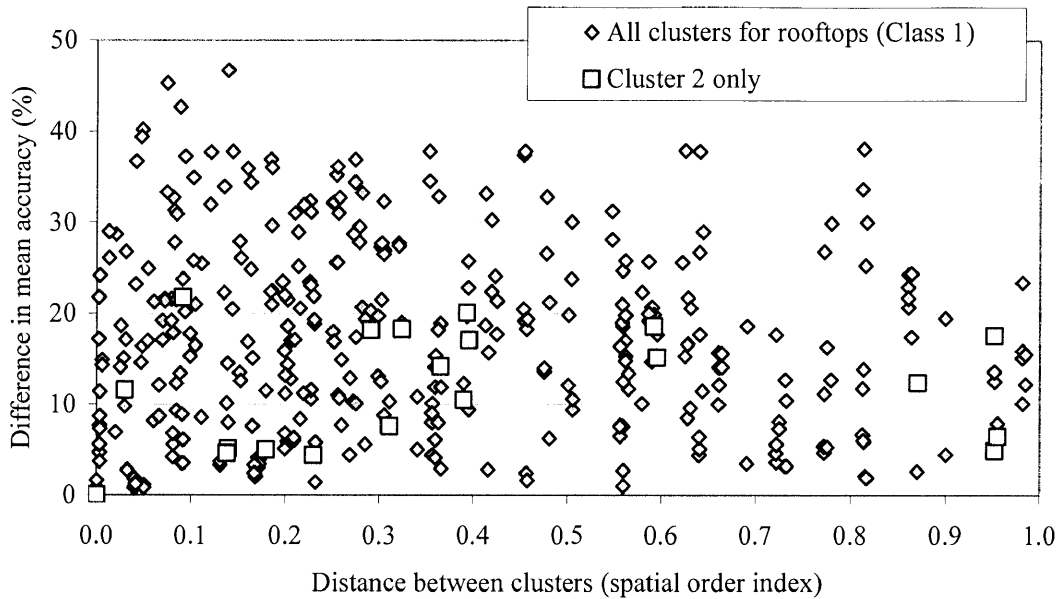
Figure 23: Difference in mean accuracy for rooftops by distance between clusters

In general, randomized pixel-level selection led to a better generalization of the classifier and improved overall classification accuracy over inputs that clustered the pixels. From a cost perspective, clusters were far more efficient for selecting input examples. This presents a tradeoff between minimizing the cost of the teacher in preparing examples and maximizing classification accuracy.

5.4 Inference

The beginning sections of this thesis presented a number of presumptions about ANN learning environments that guided the development of experiments in this thesis. The experiments substantiated some of the presumptions and did not substantiate others. For example, the presumption that training sets and test set were randomly selected from the

same distribution was enforced through the experimental design as a criterion in the selection of input examples.

The presumption that a binary classifier has a higher capacity to generalize than an n-ary classifier was not tested in this study, but an n-ary classifier may be one means to overcome problems of class separability identified in this study. It would be valuable to run similar experiments for a neural network classifier with two target concepts for cases where there was a high degree of spectral confusion, such as with roads and pavement or with low vegetation and treetops.

The presumption that increasing the sample size and number of learning cycles will result in a statistically significant reduction in the average error of prediction was explored. A change in sample size was found to have had a greater overall effect than a change in the number of epochs. There was a threshold at which neither sample size nor epochs provided a significant benefit, but that threshold was dependent on the target concept to be learned. An evaluation of overall classification accuracy indicated that fewer training examples and fewer epochs were needed in cases where there was good class separability. The results of this study did not define a rule-of-thumb that could extend to other problem domains for selecting a sample size and number of epochs. A large number of examples and a sufficient number of learning cycles will generally improve

classification accuracy, but without an evaluation framework it is difficult to determine an appropriate value for these parameters within a given problem domain.

It is a fundamental assumption that input examples are correctly classified or classified within an error tolerance acceptable to the domain objective. The learner can only be as good as the examples presented by the domain expert. Without an independent source of ground-truth data to evaluate the accuracy of JPL's classification of the Presidio image, it was difficult to determine how well a classifier might have been able to perform. If the expert, in this case JPL, inconsistently labeled the target concepts then error in representation could mask actual improvements in the learning environment. ANNs are capable of accommodating noisy input data. Nevertheless, Opitz, *et al.* (2000) indicates that overall classification accuracy can improve when an improved classification of the examples is provided to the inductive learner.

The presumption that increasing spatial context representation would result in a statistically significant reduction in classification error while retaining a high-level of generalization was shown to be generally false in these experiments. Changing the input from a single pixel to a foveal representation decreased overall classification accuracy and increased standard error. Class separability did improve in some cases using foveal representations, but the results were mixed and inconclusive.

The presumption that classification accuracy was a function of proximity to examples could not supported by this study. Classification accuracy was not a function of spatial proximity to the training set. Within in cluster variability of classification accuracy was at least great as between cluster variability for a model within a given target class. Clustering input examples resulted in a lower overall accuracy and a higher standard error than random selection of individual pixels. Interestingly, the overall classification accuracy of clusters was similar to the 5 by 5 and 9 by 9 pixel representations used by Bain (2000). As a method for selecting input examples, clusters were very efficient. A case could be made that a good generalized ANN classifier can result from examples selected from small areas of interest which are randomly selected from throughout the image or image domain.

The experiments in this study demonstrated that costs (Turney 2000) increased when the input representation changed from a single pixel to a foveal representation. The cost of misclassification, test, teacher, computation, and instability all increased with foveal representations. The clustering of inputs increased the cost of misclassification, but most other costs remained the same as the spectral classification. Clustering examples would reduce the cost of teacher in the preparation of training examples.

6. DISCUSSION AND FUTURE WORK

6.1 Discussion

Since classification is goal-oriented, it was important to the evaluation process in this study to have a means to assess tradeoffs in costs and outcomes. This investigation demonstrated that ROC curves could be efficiently produced and employed as an evaluation tool. ROC curves provided a basis to identify and potentially reject weak model components. Statistical techniques strengthened this determination. An analysis of the difference of means or the G-Statistic in combination with ROC curves offered a means to quantify differences and determine if the differences were significant. These statistics were easy and efficient to compute. Overall accuracy alone was not sufficient to fully evaluate the tradeoffs. Measures of variability were equally important to measures of classification accuracy in the evaluation and generalization of the results. Sokal and Rohlf (1981) notes that an analysis of variance is just as important as an analysis of mean accuracy. To have a measure of variation in outcomes required replication. Evaluation supported by replication could be supported in an interactive learning environment.

The experiments in this study emphasized the importance of evaluating negative and false positive cases, particularly in cases where there was evidence of spectral confusion between classes. Discovering the confusion would provide an opportunity to address class separability through transformations on existing inputs, through the addition of

other data, or through a more complex network. These opportunities need to be explored. The learner could present unclassified, negative, and false positive examples back to domain expert and ask the expert to classify these examples. This is similar to the hierarchical approach to classification investigated by Mangrich (2001). As noted by Marr and Poggio (1979), determining how much of feature identification could be learned from raw visual data optimizes identification of primitive abstractions and serve as a precursor to higher-order segmentations. There is much more opportunity to explore the advantages of pixel-level representations to classify primitive abstractions.

Setiono and Liu (1997) offer a relevant discussion on the important relationship between feature classification and feature selection. Classification is a precursor to feature selection. Feature selection is the process of identifying those inputs that are relevant and have a high predictive capacity. By selecting only the relevant attributes of the data, a higher predictive accuracy can be expected from a machine learning method. The expert may not have á priori knowledge as to which inputs are relevant. Through feature selection, attributes that are highly correlated may be minimized with little or no loss of information to the learner. There are potentially a number of domain specific transformations, like measures of texture, which could improve the number of discriminating variables in the input layer of the ANN. The experiments in this study indicate that future investigations should demonstrate the direct benefit of improving discrimination at the input level using pixel-level representations. Simple transformations on primitive inputs could potentially eliminate a large amount of the

spectral confusion between classes. ROC curves could guide the expert on what transformations to apply first or if inputs need to be filtered. Recall that only a small number of examples were required to identify confusion between classes. Partitioning the inputs, running small samples, and developing comparative ROC curves from the results could assist the domain expert in identifying strong factors that improve classification results and reduce costs. Pixel-level inputs also have the advantage of being independent of resolution, and can easily be scaled to the number of examples required for the concept to be learned. Using foveal representations at image resolutions that are lower than those in this study are likely to result in outputs that are significantly lower in overall accuracy and higher in standard error. Increasing the size or complexity of the foveal filter would increase the cost of classification.

The experiments in this study supported the feasibility of using ANNs for feature classification, and emphasized the need to evaluate the relative costs of alternative abstractions or selection strategies. The selection of inputs, input representations, and formulation of the learning environment all affected the outcome. Since any classification is done within the context of the classifier - in this case, the domain expert plus the ANN - there is a continuum of possibilities to consider when looking at target concepts and image sources. Without an evaluation framework that operates against a known baseline, it would be difficult to assess cost tradeoffs in a meaningful and quantifiable manner.

6.2: Future Work

Based on the investigations in this thesis, other questions and future work that could

address aspects of ANNs for image classification and feature identification are:

- Will a more complex neural network result in a more robust classification? A double layer of hidden nodes can represent a more complex function. Would a more complex network be a more robust classifier and improve class separability?

- How does the selection of pure versus fuzzy examples affect the outcome? Spectrally "pure" examples like shoreline are easy to classify. Would assigning membership values between 0 and 1 to inputs improve results when spectral signatures and class values are not pure representations?

- What is the impact of the selection of negative (non-class) values on the classification process? Is it important to stratify negative and positive examples in their input representation and selection?

- What are those conditions under which a foveal or filter-based classifier is better than more primitive representations?

- Is spatial context more important for some features such as roads than other features such as buildings? Is there a stronger spatial relationship between the next sections of road than the next buildings?

- Is it necessary to minimize or guide expert interaction in the learning process? Does not guiding the role of the expert introduce unforeseen bias or dependencies into the process?

- Turney (2000) identified many costs to consider, most of which are dependent on good measures of classification error and context. Can other tools be developed to assist the teacher in evaluating the relative costs associated with choices made during the learning process?

- Could semivariograms be useful in the identification and classification of linear features?

7. CONCLUSIONS

This study provided a comparative analysis of alternative input representations in an ANN learning environment, and an analysis of an alternative method of selecting inputs. Input representation and the selection of inputs did affect the probability of correctly detecting a target. Evaluating the relationship between sample size, epochs, input representation, and selection of inputs proved to be important in identifying ANN strategies that would not significantly impact the cost of correctly classifying a target. The tradeoffs in costs were minimal at the pixel-level because complexity was low, processing speed was efficient, and classification accuracy was high.

For features that were spectrally distinct from other classes, the experiments in this study supported the importance of sample size in obtaining an acceptable probability of detecting the target class. For features that were not spectrally distinct from one or more classes, neither a large sample size, nor an increased number of epochs, nor spatial context significantly improved the probability of a correct classification. The selection of examples using clusters did not generalize as well as randomly selected pixels. Classification accuracy was not dependent on proximity to the training examples. The replication of simulations provided an estimate of standard error across simulations, and a baseline for measuring model improvements. There is significant opportunity to further utilize information from replicated learning scenarios to optimize the success of the learner.

8. REFERENCES

Bain, W., (2000). MS Thesis, *A Comparison of Three Machine Learning Algorithms for Automated Feature Extraction from Digital Images*. Director: David W. Opitz, University of Montana. 48 pp.

Curran, P. (1988). The Semivariogram in Remote Sensing: An Introduction. *Remote Sensing of Environment* 24:493-507.

Drummond, C., R.C. Holte. (2000). Explicitly Representing Expected Cost: An Alternative to ROC Representation. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 198-207.

Foody, G.M., and M.K. Arora. An evaluation of some factors affecting the accuracy of classification by an artificial neural network. *Int. J. Remote Sensing*. 1997. Vol.18, no. 4, pp. 799-810.

Kanellopoulos, I, G., G. Wilkinson, F. Roli, and J. Austin (Eds.), *Neurocomputation in Remote Sensing Data Analysis*. New York: Springer-Verlag, 1997.

Langdon, W.B. (2001). Receiver Operating Characteristics (ROC). Unpublished Internet reference: http://www.cs.ucl.ac.uk/staff/W.Langdon/roc/

Mangrich, M., (2001). M.S. Thesis. *Hierarchical Feature Extraction: A Stepwise Approach to Image Classification*. Director: David W. Opitz, University of Montana. 52 pp.

Marr, D., and T. Poggio. (1979). A Computational Theory of Human Stereo Vision. *Proceedings of the Royal Society of London*, B204 (1979):3-128.

Mitchell, T., Machine Learning, Boston, MA. MIT Press, 1996. 414 pp.

Opitz, D. W. (1997). The Effective Size of a Neural Network: A Principal Component Approach, *Fourteenth International Conference on Machine Learning*, (pp. 263-271), Nashville, TN.

Opitz, D. W., M. Mangrich, J. Zeiler, and S. Blundell (2000). *A Comparison Of Feature Extraction Techniques In Remotely Sensed Imagery*. Department of Computer Science, University of Montana. 5 pp.

Platzman, L. K., and J. J. Bartholdi. Spacefilling Curves and the Planar Traveling Salesman Problem, J., *Journal of the Association of Computing Machinery*, Vol. 36, No. 4, October 1989, (pp. 719-737)

Provost, F., T. Fawcett, and R. Kohavi, (1998). The case against accuracy estimation for comparing induction algorithms. *Presented at ICML-98: Fifteenth International Conference on Machine Learning.*

Ripley, B. D. Pattern Recognition and Neural Networks. Cambridge University Press. 1996. 416 pp.

Russell, S., and P. Novig. Artificial Intelligence: A Modern Approach. Prentice-Hall. 1995, pg. 553.

Setiono, R., and H. Liu. Neural-network feature selector, *IEEE Transactions on Neural Networks*, Vol. 8, No. 3, May 1997, pages 654-662.

Sokal, R.R., and F.J. Rohlf, (1981). Biometry. San Francisco: W.H. Freeman and Company. 851 pp.

Sweet, M.D. and D.W. Opitz. 1999. A collaborative machine learning approach to object identification in digital imagery. Pages 84-89 in Weber, K.T. and S.H. Swetnam (Eds.) Proceedings of the 1999 Intermountain GIS Users Conference. 99 pp.

Swets, J.A., Measuring the accuracy of diagnostic terms. *Science* 1988; 240:1285-1293.

Turney, P.D. (1990). The Curve Fitting a Problem: A solution. *Brit. J. Phil. Sci.* 509-530.

Turney, P.D. (2000). Types of cost in inductive concept learning, *Workshop on Cost-Sensitive Learning at the Seventeenth International Conference on Machine Learning (WCSL at ICML-2000), Stanford University, California*, pages 1-7.

Wilkinson, G. G. (1997). Open Questions in Neurocomputing for Earth Observation. In: *Neurocomputation in Remote Sensing Data Analysis.* New York: Springer-Verlag. 1997.

Wilson, D. R., and T. R. Martinez. Improved Heterogeneous Distance Functions. *Journal of Artificial Intelligence Research,* 6 (1997) 1-34.