A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIS Data

Xueqiao Huang and John R. Jensen

Abstract

A machine learning approach to automated building of knowledge bases for image analysis expert systems incorporating GIS data is presented. The method uses an inductive learning algorithm to generate production rules from training data. With this method, building a knowledge base for a rule-based expert system is easier than using the conventional knowledge acquisition approach. The knowledge base built by this method was used by an expert system to perform a wetland classification of Par Pond on the Savannah River Site, South Carolina using SPOT multispectral imagery and GIS data. To evaluate the performance of the resultant knowledge base, the classification result was compared to classifications with two conventional methods. The accuracy assessment and the analysis of the resultant production rules suggest that the knowledge base built by the machine learning method was of good quality for image analysis with GIS data.

Introduction

Incorporating supplemental GIS information and human expert knowledge into digital image processing have long been acknowledged as a necessity for improving remote sensing image analysis. Enslin *et al.* (1987) pointed out that geographers should examine how GIS can be used to improve image classification through application of the logic and techniques of artificial intelligence. In recent years, a number of studies have used *expert systems* (sometimes called *knowledge-based systems*) to perform image analysis, many of which incorporate GIS data (Mckeown, 1987; Civco, 1989; Skidmore, 1989; Newkirk and Wang, 1990; Argialas and Harlow, 1990; Bolstad and Lillesand, 1992; Janssen and Middelkoop, 1992; Westmoreland and Stow, 1992; Knotoes *et al.*, 1993). The heart of the expert system approach is its *knowledge base* (Luger and Stubblefield, 1993). The usual method of acquiring knowledge in a computer-usable format to build a knowledge base involves human domain experts and knowledge engineers (Figure 1a). The domain expert explicitly expresses his or her knowledge about a subject in a language that can be understood by the knowledge engineer. The knowledge engineer translates the domain knowledge into a computer-usable format and stores it in the knowledge base.

This process presents a well-known problem when creating expert systems that is often referred to as the "knowledge acquisition bottleneck." The reasons are (Bratko, *et al.*, 1989): (1) the process requires the engagement of the domain expert and the knowledge engineer over a long period of time, and (2), although experts are capable of using their knowledge in their decision making, they are often incapable of formulating their knowledge explicitly in a form sufficiently systematic, correct, and complete to form a computer application. Some remote sensing scientists have acknowledged the difficulties in building knowledge bases for image analysis (Argialas and Harlow, 1990; Kontoes *et al.*, 1993).

To solve this problem, much effort has been exerted in the artificial intelligence community to automate knowledge acquisition to obtain low-cost and high-quality knowledge bases (Maniezzo and Morpurgo, 1993). Studies on automated knowledge acquisition belong to the subfield of artificial intelligence known as *machine learning* (Carbonell *et al.*, 1983).

Machine learning has been used to automate knowledge-base building for expert systems in many areas. Although there are some applications of machine learning techniques in the area of spatial data processing and analysis, most of them are in spatial modeling (Walker and Moore, 1988; Moor *et al.*, 1991; Aspinall, 1992). Little effort has been made to apply the techniques to automate knowledge-base building for remote sensing image analysis with GIS data. This paper describes the logic and development of a machine-learning methodology to automatically build a knowledge base for an integrated image analysis expert system that incorporates remotely sensed and GIS data. This method eliminates or reduces the difficulty caused by the "knowledge acquisition bottleneck," and should allow expert system techniques to be adopted more easily by remote sensing and GIS scientists.

Methodology

Machine learning is the science of computer modeling of learning processes. It enables a computer to acquire knowledge from existing data or theories using certain inference strategies such as induction or deduction. Over the years, research in machine learning has been pursued with varying degrees of intensity using different approaches and placing emphases on different aspects and goals (Carbonell *et al.*, 1983). In this study, we focus on one type of learning technology, *inductive learning* and its application in building knowledge bases for image analysis expert systems.

Inductive Learning

A human being has the ability to make accurate generalizations from a few scattered facts provided by a teacher or the

Department of Geography, University of South Carolina, Columbia, SC 29208.

X. Huang is now with Computer Data Systems, Inc., 1201 Elmwood Park Blvd., New Orleans, LA 70123 (xueqiao_huang@mms.gov).

Photogrammetric Engineering & Remote Sensing, Vol. 63, No. 10, October 1997, pp. 1185–1194.

0099-1112/97/6310–1185$3.00/0

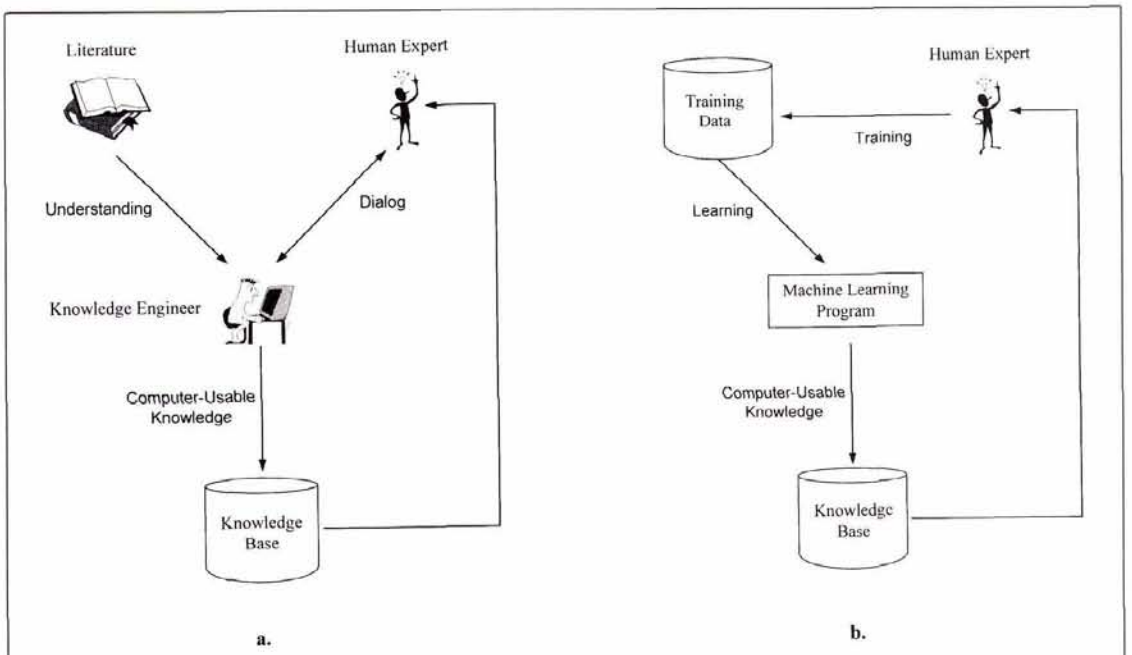

Figure 1. (a) The traditional method of knowledge acquisition used to construct a knowledge base. (b) A machine-learning approach to automated knowledge-base building.

environment using inductive inferences. This is called inductive learning (Michalski, 1983). In machine learning, the process of inductive learning can be viewed as a heuristic search through a space of symbolic descriptions for plausible general descriptions, or *concepts*, that explain the given input training data and are useful for predicting new data (Dietterich and Michalski, 1983). Inductive learning can be formulated using the following symbolic formulas (Michalski, 1983):

$$\forall i \in I \qquad (E_i \Rightarrow D_i) \tag{1}$$

$$\forall i,j \in I \qquad (E_i \Rightarrow \sim D_j), \text{ if } j \neq i \tag{2}$$

where D_i is a symbolic description of class i, E_i is a predicate that is true only for the training events of class i, I is a set of class names, $\sim$ stands for "negation," and $\Rightarrow$ stands for "implication." Expression (1) is called the *completeness condition* and states that every training event of some class must satisfy the induced description D_i of the same class. However, the opposite does not have to hold, because D_i is equivalent to or more general than E_i. This means that D_i may include some features that do not exist in some samples in E_i. Expression (2) is called the *consistency condition* and states that, if an event satisfies a description of some class, it cannot be a member of a training set of any other class. The task of inductive learning is to find through the space of descriptions the general description set $D = \{D_1, D_2, \cdots, D_i\}$ for the class set $K = \{K_1, K_2, \cdots, K_i\}$ that satisfies the completeness condition and also, in some cases, the consistency condition.

The general description set, or concept, D resulting from inductive learning can be represented by a variety of formalisms, including *production rules* (Quinlan, 1986; Quinlan, 1993). This means that inductive learning can be used to build knowledge bases for expert systems because production rules are the most popular form of knowledge representation in expert systems (Bratko, 1990; Giarratano and Riley, 1994). A motivation for the use of this approach to build a knowledge base is that it requires only a few good examples to function as training data. This is often much easier than explicitly extracting complete general theories from the domain expert (Bratko, 1990). An inductive learning approach to automated knowledge-base construction is illustrated in Figure 1b.

There are a number of inductive learning algorithms, such as Mitchell's (1982) vision spaces, Quinlan's (1986; 1993) ID3 and C4.5, and Michalski *et al.*'s (1986) AQ15. The C4.5 algorithm was selected for this research. It has the following advantages:

- The knowledge learned using C4.5 can be stored in a production rule format that can be used to create a knowledge base for a rule-based expert system.
- C4.5 is flexible. Unlike many statistical approaches, it does not depend on assumptions about the distribution of attribute values or the independence of the attributes themselves (Quinlan, 1993). This is very important when incorporating ancillary GIS data with remotely sensed data because they usually have different attribute value distributions and some of the attributes may be correlated.
- C4.5 is based on a decision-tree learning algorithm that is one of the most efficient forms of inductive learning (Bratko, 1990; Jackson, 1990). The time taken to build a decision tree increases only linearly with the size of the problem (Jackson, 1990; Quinlan, 1993).

Knowledge-Base Building Procedure

The procedure of applying the inductive learning technique to automatically build a knowledge base for a remote sensing image analysis expert system that incorporates GIS data involves training, decision tree generation, and the creation of production rules. The resultant production rules compose the knowledge base and can be used by an expert system to perform the final image classification. Figure 2 illustrates the procedure.

Training

The objective of training is to provide examples of the concepts to be learned. When building a knowledge base for image classification, the examples should be a set of training objects, each of which is represented by an attribute value-class vector such as

$$[\text{attribute}\_1, \ldots, \text{attribute}\_n, \text{class}\_i]$$

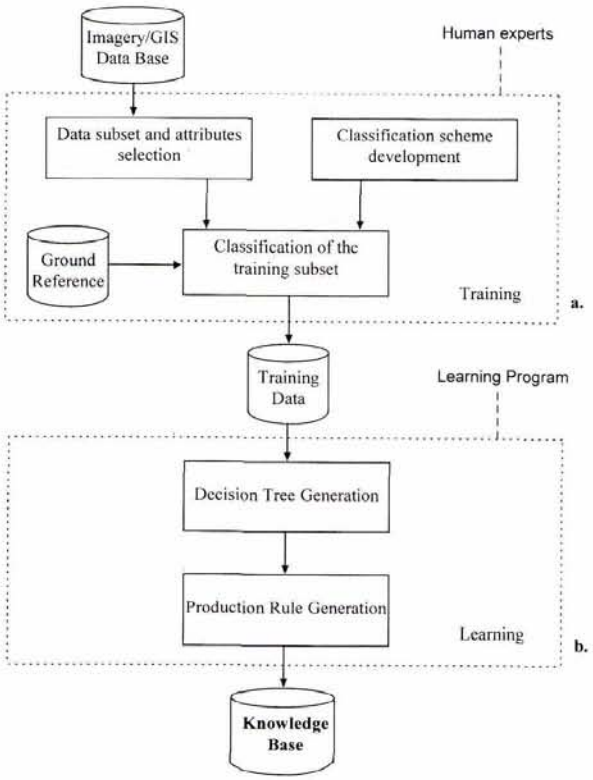

Figure 2. The procedure of using the inductive-learning technique to build a knowledge base.

The learning algorithm attempts to induce from this training data set some generalized concepts, i.e., rules that can be used to classify the remaining data. This is the process whereby a domain expert's expertise is involved. First, a subset of data is selected as training data. It should be representative of all the possible classes in the remaining unseen data. Simple random sampling may be inappropriate for this purpose because it may undersample, or even miss, small classes. Stratified random sampling is more appropriate because it guarantees that a minimum number of samples are selected from each strata (Congalton, 1988). A classification scheme must be developed at this stage. The attributes to be used in learning and classification must also be determined. The training data are then pre-classified according to the classification scheme by human experts based on their expertise and ground reference information (Figure 2a).

Decision Tree Generation

The C4.5 learning algorithm first generates decision trees from the training data. These decision trees are then transformed into production rules (Figure 2b). A decision tree can be viewed as a classifier composed of leaves that correspond to classes, decision nodes that correspond to attributes of the data being classified, and arcs that correspond to alternative values for these attributes. A hypothetical example of a decision tree is shown in Figures 3a and 3b.

A recursive "divide and conquer" strategy is used by C4.5 to generate a decision tree from a set of training data (Hunt *et al.*, 1966; Quinlan, 1993). The training data set S is divided into subsets $S_1, \ldots, S_n$ according to $a_1, \ldots, a_n$, which are the possible values of a single attribute A. This generates a decision tree with A being the root and $S_1, \ldots, S_n$ corresponding to subtrees $T_1, \ldots, T_n$ (Figure 3c). The same process is applied to the data subsets recursively to construct subtrees for each subset, until all data in a subset belong to only one class.

The stop condition for such a procedure will eventually be satisfied, resulting in a final decision tree. The goal is to build a decision tree the size of which is as small as possible. This ensures that the decision-making by the tree is efficient and effective. The goal is realized best by selecting the most "informative" attribute at each node so that it has the power to divide the data set corresponding to the node into as "pure" subsets as possible. C4.5's attribute selection criterion is based on the entropy measure from communication theory. Because entropy is in fact a measurement of impurity (Bratko, 1990), at each node, the attribute with the minimum entropy is selected to divide the data set.

From Decision Trees to Production Rules

Although the decision tree is an important form of knowledge representation, it is rarely used directly in knowledge bases in expert systems. Decision trees are often too complex to be understood, especially when they are large. A decision tree is also difficult to maintain and update. Therefore, it is often desirable to transform a decision tree to another form

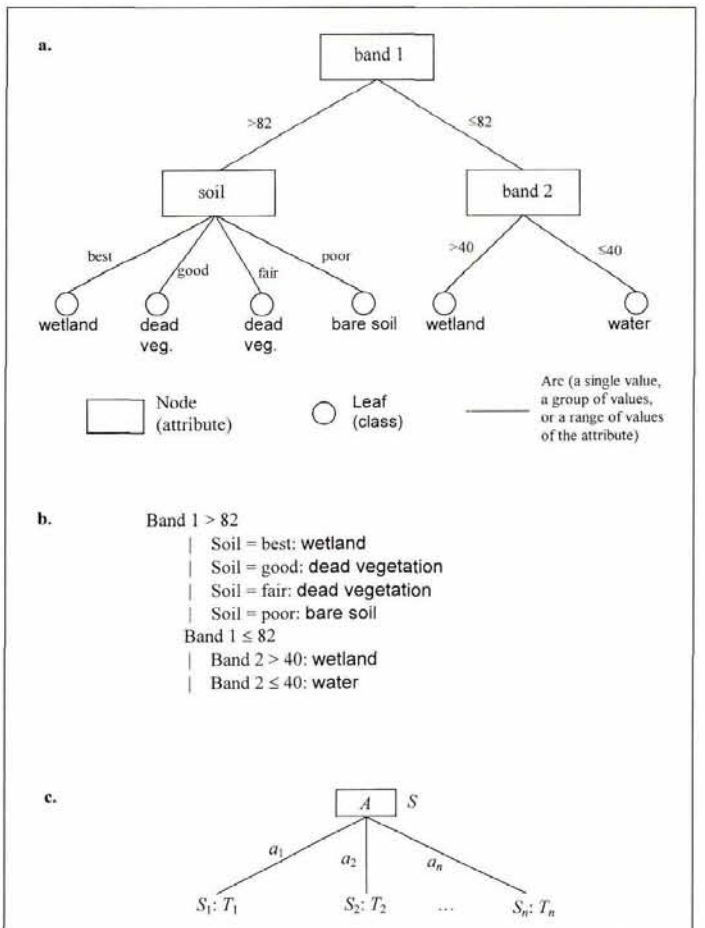

Figure 3. (a) Hypothetical example of a decision tree. A dataset consisting of three attributes (band 1, band 2, and soils) is classified into one of four classes: water, wetland, dead vegetation, and bare soil. (b) Another method of presenting the decision tree shown in (a). (c) A decision tree generated by dividing data set S into subsets.

of knowledge representation adopted commonly in expert systems, such as production rules.

A production rule can be expressed in the following general form (Jackson, 1990):

$$P_1, \ldots, P_m \rightarrow Q_1, \ldots, Q_n \quad (3)$$

with the meaning

if *premises* (or *conditions*) P_1 and ... and P_m are true, then perform *actions* Q_1 and ... and Q_n.

In fact, each path from the root to a leaf in a decision tree can be translated to a production rule. For example, the path from the root to the most left leaf in the decision tree in Figure 3a can be represented by a production rule: i.e.,

(*band 1* > 82), (*soil* = poor) → (class = *bare soil*).

There are several problems that must be solved when transforming a decision tree into production rules. First, individual rules transformed from the decision tree may contain irrelevant conditions. C4.5 uses a pessimistic estimate of the accuracy of the rule to assess a rule and decide whether a condition is irrelevant and should be deleted. Second, the rules may cease to be mutually exclusive and exhaustive. Some rules may be duplicative or may conflict. This is a common problem for rule-base building using either a manual or automated approach. Usually, a rule-based system should have some conflict resolution mechanism to deal with this problem. The approach adopted by C4.5 is ordering the sets of rules for the classes according to minimized false positive errors (the number of training objects that were incorrectly classified as class *C* by a rule set) (Quinlan, 1993). If an object can be classified into more than one class by two or more rules, the first rule that is satisfied by an object is taken as the operative one because it has the smallest possibility to assign a wrong class to the object.

Some objects in the data to be classified may satisfy no rules. This problem can be solved by defining a default rule that will fire if no other rule fires for an object. This rule in fact specifies a default class to be assigned. C4.5 uses a simple but reasonable approach: selecting as the default the class that contains the most training objects not satisfying any rule (Quinlan, 1993).

The quality of the resultant rules can be evaluated by predicting error rates derived by applying the rules on a test data set. Because the rules are easy to understand, they can also be examined by human experts. With caution, they may be edited directly.

System Implementation and Evaluation

System Implementation

An integrated system was developed to implement the proposed method. The C programming language was used for the system development on a UNIX workstation. For the purpose of testing the quality of the knowledge base built by the proposed method, the system also included an expert subsystem that used the knowledge base built by the learning subsystem to perform image classification. The components of the system are described below.

The machine learning subsystem was developed using a set of C functions provided by C4.5. The input data of the learning system was a text file with each line representing a training object. For raster remote sensing and GIS data, the training objects were represented by a pixel vector in the layer stack shown in Figure 4a. The resultant production rules were written to a file. This file became the *knowledge base* and was one of the core parts of the *expert subsystem*.

The expert subsystem was developed with the aid of an expert system shell CLIPS (Giarratano and Riley, 1994). CLIPS provided the other core part of the expert system: a forward-chaining *inference engine*. When performing classification, the knowledge base built by the learning program was first loaded into the production memory of the expert subsystem. Then, a fact converted from a pixel in the image/GIS layer stack (Figure 4a) was placed into the working memory of the expert subsystem. The inference engine then reasoned with the fact and the rules, and placed the conclusion, i.e., the classification result, into the working memory as a new fact. The data flow in the expert system is shown in Figure 4b.

The learning and expert subsystems were integrated within the ERDAS IMAGINE image processing/GIS system using the C Programmers' Toolkit and the ERDAS Macro Language (EML). The integration provided a uniform and friendly graphical user interface for the learning, expert, and image processing/GIS systems (Figure 5), and for conversion between their different data formats. It also took full advantage of other useful functions provided by the existing image processing/GIS system such as image and GIS data rectification, display, and training data selection.

System Evaluation

It was instructive to use empirical remote sensing and GIS

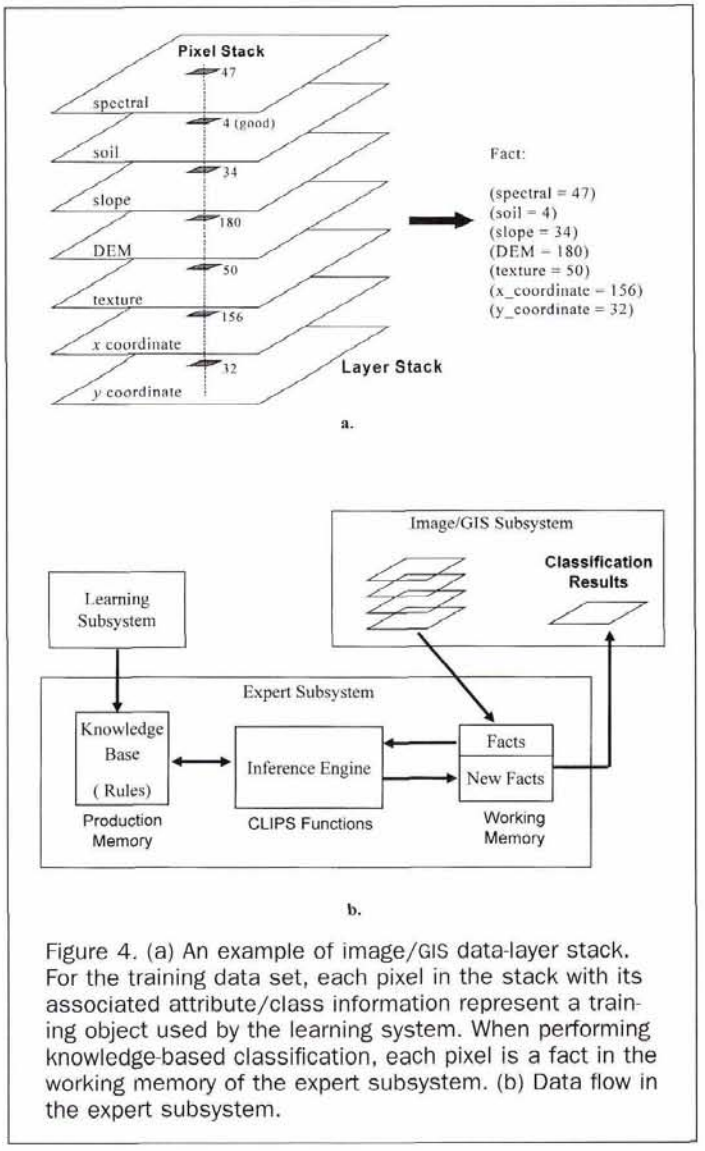

Figure 4. (a) An example of image/GIS data-layer stack. For the training data set, each pixel in the stack with its associated attribute/class information represent a training object used by the learning system. When performing knowledge-based classification, each pixel is a fact in the working memory of the expert subsystem. (b) Data flow in the expert subsystem.

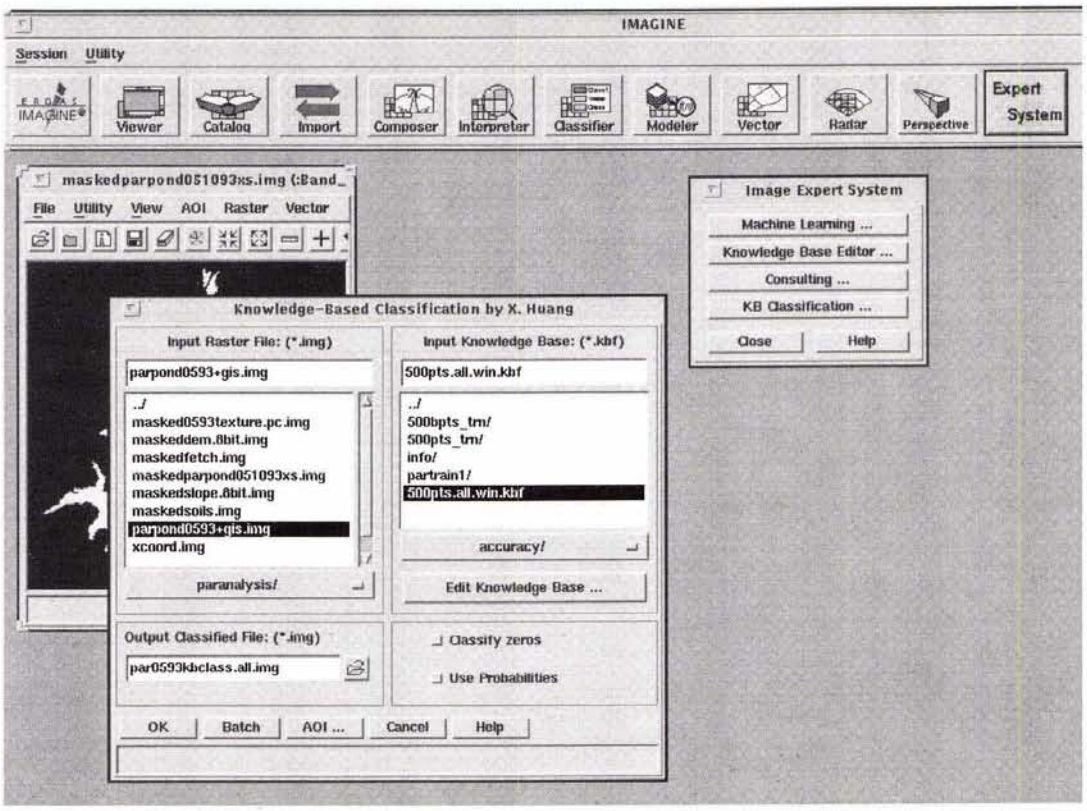

Figure 5. User interface of the machine learning and expert subsystems within ERDAS IMAGINE.

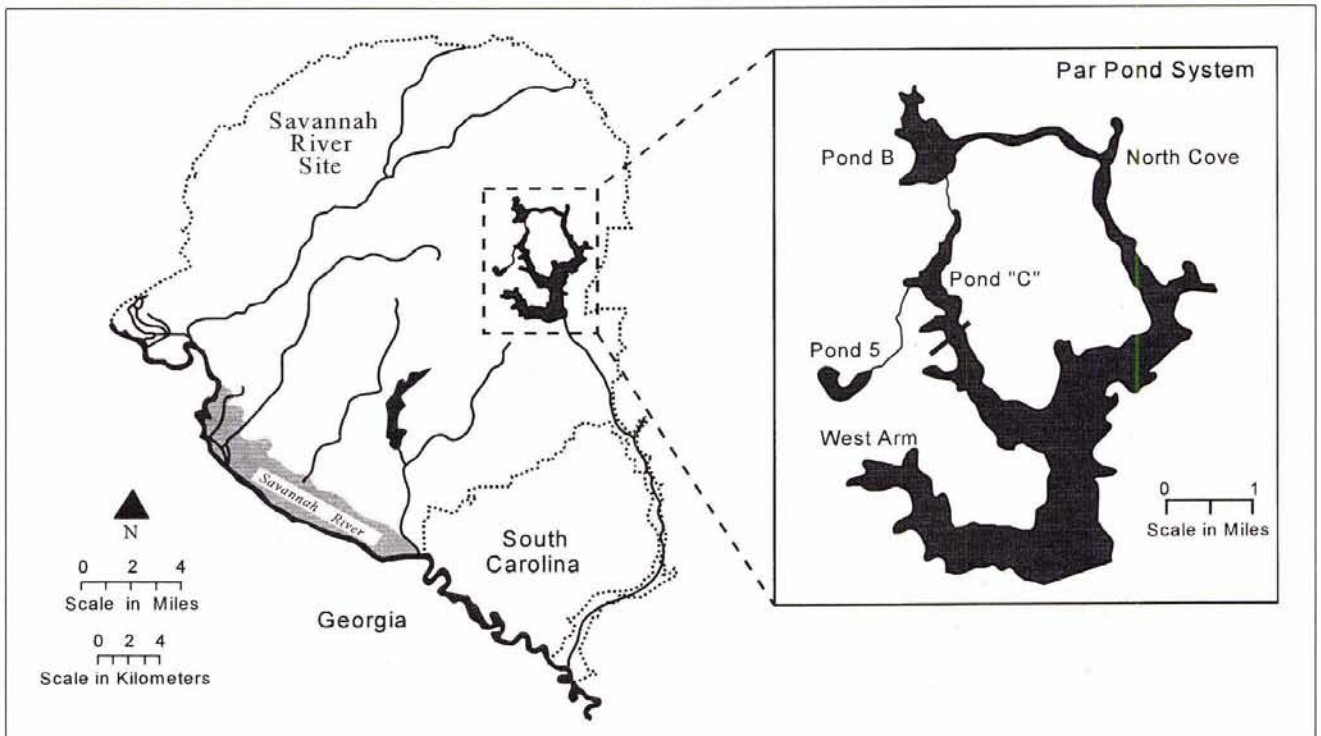

Figure 6. The Par Pond study area located on the Savannah River Site.

```
isodata in {34,45,37,44,39,42,49,35,31,38,32,25,33,40,29,46,30}:
|  dem <= 195 :
|  |  isodata in {45,46,30}: mixed
|  |  isodata in {34,37,44,39,42,49,35,31,38,32,25,33,40,29}:
|  |  |  xcoordinate > 54 : spikerush
|  |  |  xcoordinate <= 54 :
|  |  |  |  isodata in {44,49,32}: mixed
|  |  |  |  isodata in {39,42}: 6 (4.0/1.7)
|  |  |  |  isodata in {34,37,35,31,38,25,33,40,29}: spikerush
|  dem > 195 :
|  |  isodata in {34,37,35,31,38,32,25,33,29,30}: hardwood/pine
|  |  isodata in {45,39}:
|  |  |  soil = 4: dead vegetation
|  |  |  soil in {5,3,2}: mixed
|  |  |  soil = 1: dead vegetation
|  |  isodata in {44,42,49,40,46}:
|  |  |  ycoordinate <= 100 : mixed
|  |  |  ycoordinate > 100 : hardwood/pine
isodata in {1,13,3,16,5,18,2,12,4,41,47,48,43,50}:
|  dem <= 182 : water
|  dem > 182 :
|  |  isodata in {41,43}: dead vegetation
|  |  isodata in {1,13,3,16,5,18,2,12,4}: 5
|  |  isodata in {47,48,50}:
|  |  |  xcoordinate > 146 : dead vegetation
|  |  |  xcoordinate <= 146 :
|  |  |  |  fetch <= 125 : bare soil
|  |  |  |  fetch > 125 :
|  |  |  |  |  fetch <= 142 : dead vegetation
|  |  |  |  |  fetch > 142 :
|  |  |  |  |  |  fetch <= 160 : bare soil
|  |  |  |  |  |  fetch > 160 :
|  |  |  |  |  |  |  texture <= 118 : dead vegetation
|  |  |  |  |  |  |  texture > 118 : bare soil
```

a.

```
Rule 1: dem <= 182
        isodata in {1, 13, 3, 16, 5, 18, 2, 12, 4, 41, 47, 48, 43, 50}
        -> water
Rule 2: ycoordinate > 100
        isodata in {44, 42, 49, 40, 46}
        dem > 195
        -> hardwood/pine
Rule 3: fetch > 201
        xcoordinate <= 125
        isodata in {47, 48, 50}
        -> bare soil
Rule 4: fetch <= 177
        xcoordinate <= 125
        dem > 182
        isodata in {47, 48, 50}
        -> bare soil
Rule 5: texture <= 102
        xcoordinate > 125
        -> bare soil
Rule 6: isodata in {45, 46, 30}
        dem <= 195
        -> mixed
Rule 7: dem > 195
        ycoordinate <= 100
        -> mixed
Rule 8: dem <= 191
        isodata in {34, 37, 44, 39, 42, 49, 35, 31, 38, 32, 25, 33, 40,29}
        -> spikerush
Rule 9: dem > 192
        xcoordinate > 54
        -> dead vegetation

Default class: dead vegetation
```

b.

Figure 7. (a) Decision tree generated from the SPOT spectral and GIS data. (b) Production rules generated from the SPOT spectral and GIS data.

data of a freshwater reservoir in South Carolina to test whether the knowledge-base building using the proposed approach was of good quality.

The Par Pond Study Area

Par Pond is a 1000-hectare reservoir on the Savannah River Site, South Carolina (Figure 6). Natural invasion of wetland has occurred since it was constructed in 1958, with much of the shoreline having developed extensive beds of persistent and non-persistent aquatic macrophytes. Par Pond has been the object of numerous studies of wetland ecology using remote sensing and GIS techniques (Jensen *et al.*, 1992; Jensen *et al.*, 1993; Jensen *et al.*, 1997).

A SPOT multispectral (XS) image of Par Pond obtained on 10 May 1993 was used in this study. Previous studies have shown that ancillary GIS data, in addition to spectral data, are essential to the identification of some wetland vegetation. For instance, it has been confirmed that four biophysical variables [water depth or elevation, slope, fetch (unobstructed distance that wind can blow over water in a specified direction), and soils] affect aquatic macrophyte growth (Jensen *et al.*, 1992). In this study, these GIS attributes were used in conjunction with the SPOT spectral data as the initial attributes during training. All the data were rectified to a Universal Transverse Mercator map projection and resampled to 5 by 5 m. The soils data were classified into five qualitative categories according to their suitability for aquatic macrophyte growth, i.e. worst, poor, moderate, good, and best. Texture data generated from the SPOT XS data were also used. Spatial autocorrelation often exists in geographic phenomena. This suggests that objects with similar features often cluster spatially. Therefore, the spatial location of an object may be helpful to classify some geographic objects. In this research, two raster layers containing the x, y coordinates of each pixel, respectively, were used as spatial location data.

Wetland Land-Cover Classification of Par Pond

A classification scheme (Table 1) developed for a previous

TABLE 1. CLASSIFICATION SCHEME ADOPTED FOR PAR POND

Class Name	Description
Dead wetland vegetation	Dead cattails (*Typha* spp.), dead water lilies (*Nymphaea odorata*), and unknown dead vegetation on the exposed shoreline.
Spikerush	*Eleocharis quadrangulata*, a wet persistent emergent marsh.
Mixed marsh	Bulrush (*Scirpus cyperinus* (L.) Kunth) and maidencane (*Panicum hemitomon* Schult), dry persistent emergent marshes.
Old field	Grasses and forbs, usually succeed dead wetland vegetation after the submerged zone has been turned into upland for a long period.
Pine/ Hardwood	Forest on the upland surrounding the lake.
Bare soil	Bare soil.
Water	Open water of the lake.

project to monitor the successional changes in wetland land cover on the shoreline of Par Pond was adopted (Jensen *et al.*, 1997).

Two 1:20,000-scale color infrared aerial photographs obtained on 23 April 1993 were used in conjunction with *in situ* ground reference data. The aerial photos were scanned at a 300-dpi (85-μm) resolution, which corresponds to a 1.7- by 1.7-m ground resolution. The digitized aerial photographs were registered to the rectified SPOT image. A total of 550 points were selected using stratified random sampling. These points were superimposed on the screen on top of the rectified aerial images. The land-cover class at each point was determined by ecologist experts who have been working on the wetland ecology of the Savannah River Site (SRS) for more than 20 years. A total of 121 points were discarded due to the uncertainty in the class interpretation. The final 429 points were then split into two data sets; one consisting of 220 points that were only used for training and the other with 209 points that were only used for accuracy assessment.

The three-band SPOT XS data were first pre-classified into 50 spectrally homogeneous classes using the ISODATA clustering algorithm before they were used for learning and classification using the proposed approach. Although this was not required by the learning algorithm, it reduced the dimension of the spectral data from three to one. The spectral class layer was then integrated with the other six GIS data layers to form the layer stack illustrated in Figure 4a. The procedures illustrated in Figure 2 and Figure 4b were applied to this layer stack to perform machine learning and classification.

To evaluate the quality of the knowledge base built by the machine-learning approach, two conventional classifications were also performed and compared with the classification performed by the expert system with the machine-learning-derived knowledge base:

- *Traditional supervised classification with spectral and GIS data.* A maximum-likelihood algorithm was used. Three SPOT bands and six GIS data layers were analyzed. This is the "logical channel" classification approach described in Hutchinson (1982).
- *Traditional unsupervised classification with only spectral data.* The standard statistical unsupervised approach was used (Jensen, 1996). Fifty spectrally homogeneous clusters were generated using the ISODATA clustering algorithm. These clusters were combined and labeled by the experts into the six land-cover classes.

Results and Discussion

The decision trees and production rules generated by the machine-learning-assisted expert-system approach are shown in Figures 7a and 7b, respectively. Tables 2 through 5 summarize the accuracy and *KAPPA* statistics associated with each type of classification and the *z* values from the comparisons between the different classifications.

With an overall accuracy of 74.16 percent and K_{hat}

TABLE 2. ACCURACY ASSESSMENT OF A CLASSIFICATION DERIVED FROM THE MACHINE-LEARNING-ASSISTED EXPERT SYSTEM

	Ground Reference						
Classification	Water	Dead Vegetation	Spikerush	Mixed	Hardwood	Bare Soil	User's Accuracy
Water	27	1	0	0	0	0	96.42
Dead Vegetation	0	31	2	4	1	6	70.45
Spikerush	1	1	31	7	1	2	72.09
Mixed	1	6	5	19	2	0	57.58
Hardwood	0	1	3	1	24	0	82.76
Bare Soil	0	9	0	0	0	23	71.88
Producer's Accuracy	93.1	63.26	64.02	61.26	85.71	82.14	Overall 74.16

$K_{hat} = 0.68757$, $V(K) = 0.001358$

TABLE 3. ACCURACY ASSESSMENT OF A CLASSIFICATION DERIVED FROM A MAXIMUM-LIKELIHOOD ANALYSIS USING BOTH SPECTRAL AND GIS DATA

	Ground Reference						
Classification	Water	Dead Vegetation	Spikerush	Mixed	Hardwood	Bare Soil	User's Accuracy
Water	26	0	0	0	0	0	100.00
Dead Vegetation	0	25	1	3	5	13	53.19
Spikerush	2	0	27	7	0	0	75.00
Mixed	1	12	10	20	3	1	43.47
Hardwood	0	1	1	1	20	0	86.96
Bare Soil	0	11	2	0	0	18	58.06
Producer's Accuracy	89.66	51.02	65.86	64.52	71.42	58.06	Overall 65.07

$K_{hat} = 0.57757$, $V(K) = 0.001613$

TABLE 4. ACCURACY ASSESSMENT OF A CLASSIFICATION DERIVED FROM AN UNSUPERVISED ISODATA ALGORITHM USING ONLY SPECTRAL DATA

	Ground Reference						
Classification	Water	Dead Vegetation	Spikerush	Mixed	Hardwood	Bare Soil	User's Accuracy
Water	27	0	1	0	0	0	96.42
Dead Vegetation	0	18	0	2	1	1	81.82
Spikerush	1	4	29	11	16	0	47.54
Mixed	1	5	11	16	1	1	45.71
Hardwood	0	0	0	2	10	1	82.76
Bare Soil	0	22	0	0	0	28	56.00
Producer's Accuracy	93.1	36.73	70.73	51.61	35.71	90.3	Overall 61.24

$K_{hat} = 0.53352$, $V(K) = 0.001589$

TABLE 5. RESULTS OF Z TESTS FOR THE ERROR MATRICES OF THE THREE CLASSIFICATION APPROACHES

	Maximum-Likelihood	Unsupervised
Machine-learning-assisted expert system with spectral and GIS data	2.012 (S)	2.838 (SS)
Maximum-likelihood with spectral and GIS data		0.776 (NS)

NS – Difference is not significant at 0.95 confidence level ($Z < z_{.025} = 1.960$).
S – Difference is significant at 0.95 confidence level ($Z \geq z_{.025} = 1.960$).
SS – Difference is significant at 0.99 confidence level ($Z \geq z_{.005} = 2.575$).

= 0.6876, the proposed machine-learning approach yielded the highest accuracy (Table 2). The Z tests reveal that this approach was significantly different from the two other approaches at the 95 percent confidence level (Table 5).

While the accuracy of the maximum-likelihood with spectral and GIS data approach (Table 3) was slightly higher than that from the unsupervised approach with only spectral data (Table 4), the Z tests revealed that these two approaches were not significantly different (Table 5). On the other hand, the proposed approach was significantly different from the maximum-likelihood with spectral and GIS data approach.

The reason why the performance of the maximum-likelihood with spectral and GIS data approach was not good may be the distribution of the incorporated GIS data. An important assumption in the maximum-likelihood classification is that the data distribution for each class is Gaussian (normally distributed). However, this assumption is commonly not valid for ancillary GIS data (Hutchinson, 1982). GIS data often have a bi- or multimodal distribution. Therefore, the maximum-likelihood classification is not appropriate for such data. On the other hand, the rule-based approach does not have such a data distribution requirement. This is demonstrated in Figure 8 where bare soil (B) and dead vegetation (D) training data are displayed in SPOT XS band 2-band 3 feature space and fetch-ISODATA feature space. The maximum-likelihood classifier performs very well when processing the data that are normally distributed as suggested in Figure 8a. However, when the data have a bi- or multimodal distribution, as shown in Figure 8b, the maximum-likelihood classifier should perform poorly. Unfortunately, this is often the case when GIS data are incorporated. For example, consider production rules 3 and 4 (Figure 7b) generated from the machine learning approach which indicate that, in a two-dimensional feature space with fetch and ISODATA being the x, y axes, respectively, bare soil has a bimodal distribution. One cluster is located at the area with x (fetch) > 201, while the other cluster at the area with $x \leq 177$. This is approximately the situation in Figure 8b. The rule-based approach can handle this situation better by applying several production rules (like 3 and 4) to deal with the two bare soil clusters separately.

The error matrices indicate that there was substantial confusion between dead vegetation and bare soil in the other two classifications used for comparison. In fact, the ISODATA clusters 47, 48, and 50 were found to be the mixed classes of dead vegetation and bare soil during the cluster labeling with the unsupervised classification approach. Therefore, spectral data alone were not capable of distinguishing these two classes.

The machine-learning approach obtained significant improvements for these two classes (Table 6). From the decision tree and production rules (Figure 7) generated from this approach, it is obvious that the GIS data played an important role in the improvements. For example, fetch, DEM, and the x-coordinate were used to distinguish these two problematic classes in the following rules:

Rule 3: (isodata ∈ {47,48,50}), (fetch > 201), (xcoordinate ≤ 125) → (class = bare soil)
Rule 4: (isodata ∈ {47,48,50}), (fetch ≤ 177), (xcoordinate ≤ 125), (DEM>172) → (class = bare soil)

If a pixel's spectral (isodata) cluster value was 47, 48, or 50, but its GIS attributes did not satisfy the conditions in either rule above, this pixel was assigned to dead vegetation rather than bare soil as dead vegetation is the default class.

The machine-learning approach also revealed its "intelligence" when classifying hardwood/pine. Hardwood/pine is often spectrally confused with wetland vegetation (spikerush or mixed wetland vegetation). As hardwood/pine belongs to upland vegetation, elevation should be a useful attribute to distinguish it from wetland vegetation. The machine-learning approach discovered such a rule. Rule 2 in Figure 7b indicated that hardwood/pine was mostly distributed at elevations greater than 195 m. Therefore, DEM became an important attribute to distinguish hardwood/pine from wetland vegetation.

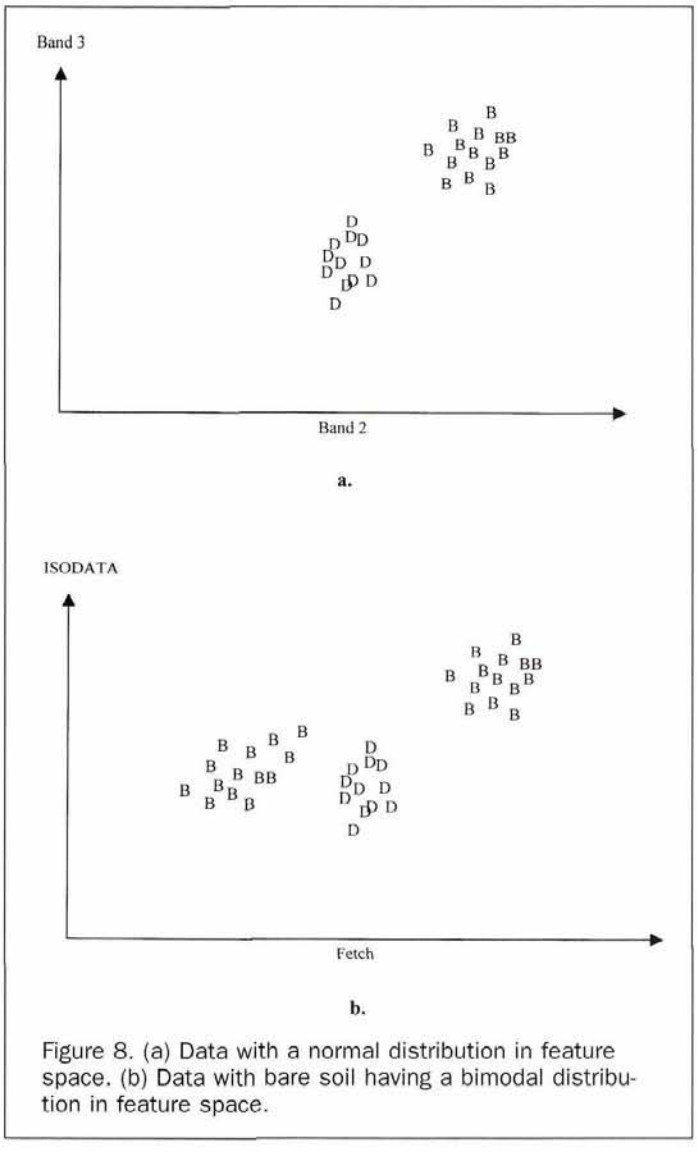

Figure 8. (a) Data with a normal distribution in feature space. (b) Data with bare soil having a bimodal distribution in feature space.

TABLE 6. COMPARISON OF THE CLASSIFICATION ACCURACY OF DEAD VEGETATION AND BARE SOIL USING THREE CLASSIFICATION ALTERNATIVES

a. Machine-Learning-Assisted Expert System Using Spectral and GIS Data

	Dead Vegetation	Bare Soil	User's Accuracy
Dead Vegetation	31	6	70.45
Bare Soil	9	23	71.88
Producer's Accuracy	63.26	82.14	Overall 78.3

b. Maximum-Likelihood with Spectral and GIS Data

	Dead Vegetation	Bare Soil	User's Accuracy
Dead Vegetation	25	13	53.19
Bare Soil	11	18	58.06
Producer's Accuracy	51.02	58.06	Overall 64.2

c. Unsupervised Classification

	Dead Vegetation	Bare Soil	User's Accuracy
Dead Vegetation	18	1	81.82
Bare Soil	22	28	56.00
Producer's Accuracy	36.73	90.32	Overall 66.6

Even though the proposed method resulted in the best classification accuracy among the three methods, the 74.16 percent overall accuracy is still relatively low. There are several reasons for this result. First, the major classes of mixed marsh, dead wetland vegetation, and spikerush represent Level III classes in the USGS Land Use/Land Cover Classification System (Anderson *et al.*, 1976; USGS, 1992). Because the USGS classification Level II usually requires remote sensor data with resolution equal to SPOT panchromatic data (Jensen, 1996), the relatively low resolution (20 by 20 m) of the SPOT XS data used for classification may be a factor. Second, the aquatic macrophytes grow like a belt along the shoreline of Par Pond. In some places, the belt is very narrow, with the width being smaller than the resolution of the SPOT XS data. This produces pixels with mixed information classes and may cause classification errors. Studies on inland aquatic macrophytes in this area usually only obtain accuracies from 65 to 70 percent due to the complex heterogeneity of materials within the IFOV of the sensor (Hodgson *et al.*, 1987; Jensen *et al.*, 1993; Jensen *et al.*, 1997).

Conclusions

A method of automated knowledge-base construction for image analysis expert systems with GIS data was developed based on an inductive machine-learning technique. With this method, building a knowledge base for a rule-based expert system for remote sensing image analysis with GIS data is easier than using the conventional knowledge acquisition approach. It does not require that domain experts explicitly express their knowledge and does not require knowledge engineers to code the knowledge. However, it is imperative that appropriate training data be selected. An operational image expert system was developed to test the utility of the knowledge base generated by the machine-learning approach.

The accuracy assessment and the analysis of the resultant production rules suggest that the knowledge base built by the machine-learning method was of good quality for image analysis. With several types of GIS data, it produced results superior to those of conventional approaches. This study demonstrated the utility of GIS data to improve remote sensing classification. It also demonstrated that selecting the appropriate approach when incorporating GIS data was very important. Because GIS data usually do not meet the Gaussian distribution assumption, maximum-likelihood classification may not be an appropriate method. On the other hand, the expert-system approach proved to be a robust and effective way to incorporate GIS data because it does not have such a data distribution requirement.

The research also demonstrated some other advantages of the machine-learning-assisted expert-system approach: it was easy to understand, and the resultant knowledge could be used in other applications. For example, from Rule 2 in Figure 7b, one can know that most of the hardwood/pine class was distributed above 195 m in the Par Pond area. Such spatial knowledge is very useful in many geographic applications such as spatial analysis and modeling. However, such data cannot be obtained from the conventional statistical classifications using the maximum-likelihood classification algorithm.

Acknowledgments

We would like to thank Drs. D. Cowen, J. Rose, and D. Wagner for their comments and suggestions, and Dr. H. Mackey, Jr., for providing the data and expertise on the Savannah River Site ecology.

References

Anderson, J.R., E. Hardy, J. Roach, and R. Witmer, 1976. *A Land Use and Land Cover Classfication System for Use with Remote Sensing Data*, U.S. Geological Survey Profession Paper 964, Washington, D.C., 28 p.

Argialas, D., and C. Harlow, 1990. Computational image interpretation models: An overview and perspective, *Photogrammetric Engineering & Remote Sensing*, 56(6):871–886.

Aspinall, R., 1992. An inductive modelling procedure based on Bayes' theorem for analysis of pattern in spatial data, *Int. J. Geographical Information Systems*, 6(2):105–121.

Bolstad, P.V., and T.M. Lillesand, 1992. Rule-based classification models: Flexible integration of satellite imagery and thematic spatial data, *Photogrammetric Engineering & Remote Sensing*, 58(7):965–971.

Bratko, I., 1990. *PROLOG: Programming for Artificial Intelligence, Second Edition*, Addison-Welsey Publishing, Wokingham, England, 597 p.

Bratko, I., I. Kononenko, N. Lavrac, I. Mozetic, and E. Roskar, 1989. Automatic synthesis of knowledge: Ljubljana research, *Machine and Human Learning* (Y. Kodratoff and A. Hutchinson, editors), GP Publishing, Inc., Columbia, Maryland, pp. 25–33.

Carbonell, J.G., R.S. Michalski, and T.M. Mitchell, 1983. An overview of machine learning, *Machine Learning, Vol. I* (R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors), Morgan Kaufmann, Inc., San Mateo, California.

Civco, D.L., 1989. Knowledge-based land use and land cover mapping, *Proc. ASPRS/ACSM Annual Convention*, Baltimore, Maryland, pp. 276–289.

Congalton, R.G., 1988. A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data, *Photogrammetric Engineering & Remote Sensing*, 54(5):593–600.

Dietterich, T.G., and R.S. Michalski, 1983. A comparative review of selected methods for learning from examples, *Machine Learning, Vol. 1* (R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors), Morgan Kaufmann, Inc., San Mateo, California.

Enslin, W.R., J. Tonand, and A. Jain, 1987. Land cover change detection using a GIS-guided feature-based classification of Landsat Thematic Mapper data, *Proc. ASPRS*, 6:108–20.

Giarratano, J., and G. Riley, 1994. *Expert Systems: Principles and Programming, Second Edition*, PWS Publishing, Boston, Massachusetts.

Hodgson, M.E., J.R. Jensen, H.E. Mackey, and M.C. Coulter, 1987. Remote sensing of wetland habitat: A wood stork example, *Photogrammetric Engineering & Remote Sensing*, 53(8):1075–1080.

Hunt, E.B., J. Marin, and P.J. Stone, 1966. *Experiments in Induction*, Academic Press, New York.

Hutchinson, C.R., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement, *Photogrammetric Engineering & Remote Sensing*, 48(1):123–130.

Jackson, P., 1990. *Introduction to Expert Systems, Second Edition*, Addison-Wesley Publishing Company, Wokingham, England.

Janssen, L.L.F., and H. Middelkoop, 1992. Knowledge-based crop classification of a Landsat Thematic Mapper image, *Int. J. of Remote Sensing*, 13(15):2827–2837.

Jensen, J.R., 1996. *Introductory Digital Image Processing: A Remote Sensing Perspective, Second Edition*, Prentice-Hall, Englewood Cliffs, New Jersey.

Jensen, J.R., S. Narumalani, O. Weatherbee, K.S. Morris, and H.E. Mackey, 1992. Predictive modeling of cattail and waterlily distribution in a South Carolina reservoir using GIS, *Photogrammetric Engineering & Remote Sensing*, 58(11):1561–1568.

Jensen, J.R., S. Narumalani, O. Weatherbee, and H.E. Mackey, 1993. Measurement of seasonal and yearly cattail and waterlily changes using multidate SPOT panchromatic data, *Photogrammetric Engineering & Remote Sensing*, 59(4):519–525.

Jensen, J.R., X. Huang, and H.E. Mackey, 1997. Remote sensing of successional changes in wetland vegetation as monitored during a four-year drawdown of a former cooling lake, *Applied Geographic Studies*, 1(1):31–44.

Kontoes, C., G. Wilkingson, A. Burrill, S. Goffredo, and J. Megier, 1993. An experimental system form the integration of GIS data in knowledge-based image analysis for remote sensing of agriculture, *Int. J. Geographical Information Systems*, 7(3):247–262.

Luger, G.F., and W.A. Stubblefield, 1993. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving, Second Edition*, The Benjamin/Cummings Publishing Company, Inc., Redwood City, California, 740 p.

McKeown, D.M., 1987. The role of artificial intelligence in the integration of remotely sensed data with geographic information systems, *IEEE Transactions on Geoscience and Remote Sensing*, 25(3):330–348.

Maniezzo, V., and R. Morpurgo, 1993. D-KAT: A deep knowledge acquisition tool, *Expert Systems*, 10(3):157–165.

Michalski, R.S., 1983. A theory and methodology of inductive learning, *Machine Learning, Vol. I* (R.S. Michalski, J.G. Carbonell, and T.M. Mitchell, editors), Morgan Kaufmann Publishers, Inc., San Mateo, California.

Michalski, R.S., J. Hong, I. Mozetic, and N. Lavrac, 1986. The multipurpose incremental learning system AQ15 and its testing application to three medical domains, *Proceedings of the Fifth Annual National Conference on Artificial Intelligence*, Philadelphia, pp. 1041–1045.

Mitchell, T., 1982. Generalization as search, *Artificial Intelligence*, 18:203–226.

Moore, D.M., B.G. Lees, and S.M. Davey, 1991. A new method for predicting vegetation distributions using decision tree analysis in a geographic information system, *Environmental Management*, 15(1):59–71.

Newkirk, R.T., and F. Wang, 1990. A common knowledge database for remote sensing and geographic information in a change-detection expert system, *Environment and Planning (B)*, 17(4):395–404.

Quinlan, J.R., 1986. Induction of decision tree, *Machine Learning*, 1(1):81–106.

———, 1993. *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, San Mateo, California.

USGS, 1992. *Standards for Digital Line Graphs for Land Use and Land Cover Technical Instructions*, Referral STO-1-2, U.S. Government Printing Office, Washington, D.C., 60 p.

Walker, P.A., and D.M. Moore, 1988. SIMPLE: An inductive modelling and mapping tool for spatially-oriented data, *Int. J. Geographical Information Systems*, 2(4):347–363.

Westmoreland, S., and D.A. Stow, 1992. Category identification of changed land-use polygons in an integrated image processing geographic information system, *Photogrammetric Engineering & Remote Sensing*, 58(11):1593–1599.