A Machine Learning Approach for Knowledge Base Construction Incorporating GIS Data for Land Cover Classification of Landsat ETM+ Image

Hwahwan Kim* · Cha Yong Ku**

지식 기반 시스템에서 GIS 자료를 활용하기 위한 기계 학습 기법에 관한 연구 - Landsat ETM+ 영상의 토지 피복 분류를 사례로

김화환* · 구자용**

Abstract : Integration of GIS data and human expert knowledge into digital image processing has long been acknowledged as a necessity to improve remote sensing image analysis. We propose inductive machine learning algorithm for GIS data integration and rule-based classification method for land cover classification. Proposed method is tested with a land cover classification of a Landsat ETM+ multispectral image and GIS data layers including elevation, aspect, slope, distance to water bodies, distance to road network, and population density. Decision trees and production rules for land cover classification are generated by C5.0 inductive machine learning algorithm with 350 stratified random point samples. Production rules are used for land cover classification integrated with unsupervised ISODATA classification. Result shows that GIS data layers such as elevation, distance to water bodies and population density can be effectively integrated for rule-based image classification. Intuitive production rules generated by inductive machine learning are easy to understand. Proposed method demonstrates how various GIS data layers can be integrated with remotely sensed imagery in a framework of knowledge base construction to improve land cover classification.

Key Words : machine learning, knowledge base, rule-based classification, land cover classification

요약 : 원격탐사에서 위성 영상의 디지털 처리 기술이 발달하면서 GIS 자료와 지식 기반 전문가 시스템과의 통합에 대한 관심이 증가하고 있다. 본 연구에서는 위성영상을 토지피복 분류하는 과정에서 GIS 자료를 통합하기 위하여 기계 학습 기법과 규칙 기반 분류 기법을 적용하였다. 사례 지역을 대상으로 Landsat ETM+ 영상과 고도, 경사, 향, 수역과의 거리, 도로와의 거리, 인구밀도 등의 GIS 자료를 함께 활용하였다. C5.0 추론 기계 학습 알고리듬을 이용하여 350개의 표본점으로부터 결정 트리와 분류 규칙을 생성하였다. 본 연구에서 도출된 규칙을 이용하여 분류한 결과, 고도, 수역과의 거리, 인구밀도 등의 GIS 자료가 규칙 기반 분류에 효과적인 것으로 나타났다. 본 연구에서 제안한 기계 학습과 지식 기반 분류 기법을 이용하면 다양한 GIS 자료들을 통합하여 위성영상을 보다 효과적으로 분류할 수 있다.

주요어 : 기계 학습, 지식 기반 시스템, 규칙 기반 분류, 토지 피복 분류

---

* Ph D. Candidate, Department of Geography, University of Georgia, khh008@uga.edu
** Associate Professor, Department of Geography, Sangmyung University, koostar@smu.ac.kr

1. Introduction

Advances in census, ground survey, and remote sensing (RS) have enabled collection of huge amount of data in large databases. Explosively growing volume of data creates the necessity of knowledge discovery from data, which leads to a promising emerging field, called data mining or knowledge discovery in databases. Remotely sensed data are the major source of the data which are continuously acquired and stored. Incorporating supplementary geographic information system (GIS) data 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. A number of studies have used expert systems (knowledge based systems) to perform image analysis, many of which incorporated GIS data (Westmoreland and Stow, 1992; Knotoes *et al.*, 1993; Huang and Jensen, 1997). The heart of expert system is in its knowledge base. The usual method for construction of knowledge base involves human domain experts and knowledge engineers who translate the domain knowledge into a computer-recognized format and store it in the knowledge base. This process presents a well-known problem that is often referred to as the "knowledge acquisition bottleneck." The reasons are: (1) the process requires the engagement of the domain expert and 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, and finally (3) rapidly increasing volume of data acquired daily poses another challenges against effective updates of the knowledge base.

Spatial data mining and knowledge discovery is the extraction of implicit, interesting spatial or non-spatial patterns and general characteristics from a large volume of digitally stored database. It provides a new way of knowledge acquisition for image classification. In this emerging research field, much effort has been exerted in the artificial intelligence community to automate knowledge acquisition to obtain low-cost and high-quality knowledge base. Studies on automated knowledge acquisition belong to the subfield of artificial intelligence known as machine learning. Eklund *et al.* (1998) extracted knowledge from TM images and geographic data in soil salinity analysis using inductive learning algorithm C4.5. Huang *et al.* (1997) extracted knowledge from GIS data and SPOT multispectral image in wetland classification using C4.5. In both studies, GIS data were converted to raster format in which the sampling size is equal to image pixel size. Qi and Zhu (2003) applied an inductive learning algorithm to extract knowledge of soil-landscape models from a soil map.

In this research, possibilities of applying inductive machine learning for remotely sensed image classification in terms of knowledge discovery are explored. Major research issues in inductive machine learning and its utility in remote sensing applications are reviewed first for introduction. Then, an inductive machine learning algorithm as alternative method of knowledge construction for expert system classification to incorporated GIS data is proposed. In order to verify the feasibility of the inductive learning based knowledge construction, a land cover classification experiment is performed with a Landsat ETM+ multispectral image and GIS datasets. Rules discovered by inductive learning

using GIS data are used to post-process the result of unsupervised ISODATA (Iterative Self-Organizing Data Analysis Technique) classification.

2. Inductive machine learning

Over the past decades, researchers in computer science have developed a number of techniques aimed at pattern recognition, classification, cluster analysis, prediction and simulation that offer huge potential when applied to geographic data analysis (Malebra *et al.*, 2001; Gahegan, 2003). Neural networks and self-organizing map, genetic algorithms, and decision trees are examples of such techniques. Inductive machine learning is one of their common properties. Machine learning is a subfield of artificial intelligence (AI), mimicking human learning processes by computer modeling. One of its major objectives is to automate the process of knowledge acquisition for other AI applications including expert systems. Inductive machine learning offers a means to characterize a category or function without relying on a priori knowledge. Inductive learning tools are trained to recognize pattern or to predict outcomes by generalizing from a group of measurements for which the outcome is known (training data) to a larger data set.

The development of inductive learning tools is driven by the need to address a range of complex, non-deterministic problems, where a truly optimal solution becomes computationally intractable (Gahegan, 2000). The need for an approximate solution may be as a result of complex input data, complex output characteristics, or a combination of both. The vast increase in the number of geospatial data sources such as multispectral remote sensing images, digital census data, and real time data from mobile devices provide a wealth of data that might be used in geographic analysis and pose considerable computational challenges. Established techniques such as maximum likelihood classification, principal component analysis, and statistical pattern recognition techniques, may be less able to address this complexity. Inductive learning tools offer significant advances over traditional techniques in that they are robust in the presence of noise, able to work with data of high dimensionality, and make fewer prior assumptions about data distribution and model parameters.

Inductive machine learning can be effectively utilized to automatically construct rule bases for expert or knowledge base systems. Inductive machine learning programs provide both an improvement on interview-based acquisition techniques and a basis for data mining methods (Quinlan, 1986). In selective domains, inductive learning systems are able to determine decision rules by inductive inference from examples of expert decisions and there are many successful generic systems that employ these techniques (Michalski and Chilausky, 1980; Quinlan, 1986). These systems demonstrated a dramatic increase in the speed of rule base construction as well as the capacity to transfer knowledge from domain expert to machine.

Advances in satellite technology and availability of downloaded images constantly increase the sizes of remote sensing image archives. Automatic content extraction, classification and content-based retrieval have become highly desired goals for the development of intelligent remote sensing databases. The common approach for mining these databases uses rules created by analysts. However, incorporating GIS data and human expert knowledge with digital image processing improves remote sensing image analysis. In this study, we develop a system that uses decision tree classifiers for interactive

learning of land cover models and mining of image archives. Decision trees provide a promising solution for this problem because they can operate on both numerical (continuous) and categorical (discrete) data sources, and they do not require any assumptions about neither the distributions nor the independence of attribute values. This is especially important for the fusion of measurements from different sources like spectral data, digital elevation model (DEM) data and other ancillary GIS data. Furthermore, using surrogate splits provides the capability of dealing with missing data during both training and classification, and enables handling instrument malfunctions or the cases where one or more measurements do not exist for some locations.

3. Methodology

Machine learning 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. This study focused on one type of learning technology, inductive learning and its application in building knowledge bases for image classification system especially combined with ISODATA unsupervised classification. Figure 1 illustrates the knowledge construction using inductive machine learning technique. Sample dataset for machine learning is created as training dataset using ancillary GIS data, spectral clusters which are acquired by unsupervised ISODATA classification, and reference classified land cover data. The knowledge base discovered by C5.0 algorithm is a group of classification rules and a default class.

1) Unsupervised classification
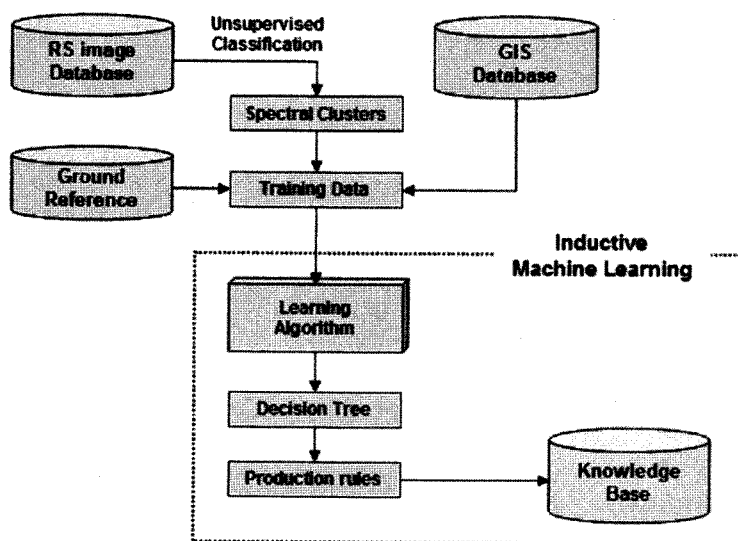
Unsupervised classifications are used frequently

Figure 1. Knowledge base construction procedure using inductive machine learning

in remote sensing analyses. Pixels with similar spectral reflective characteristics are grouped into distinct cluster. These spectral clusters are then labeled with a certain class name. The important difference to supervised classification is that classes do not have to be defined a priori. For a successful supervised classification all classes have to be known and the spectral properties of these classes have to be derived (usually through a training stage). Since unsupervised classification groups pixels into spectral clusters it is possible to discover classes that were not known before the classification.

2) Inductive machine learning using C5 algorithm

Human being has the ability to make generalizations from a few scattered facts provided by environment using inductive inference. This is called inductive learning. In machine learning, the process of inductive learning can be viewed as a search for heuristic rules between predictive and dependent variables from given training data. There are a number of inductive learning algorithms, such as Mitchell's vision spaces (Mitchell, 1997), Quinlan's ID3 (Quinlan, 1986) and C5.0 etc. The C5.0 algorithm was selected for this research. It has the following advantages:
  - The knowledge learned using C5.0 can be stored in a production rule format that can be used to create a knowledge base for rule-based expert system.
  - C5.0 is flexible. Unlike many statistical approaches, it does not depend on assumptions about the distribution of attribute. 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.

  - C5.0 is based on a decision tree learning algorithm that is one of the most efficient forms of inductive learning.

3) Knowledge base construction

The application of inductive learning technique for knowledge base construction 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.

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 a vector of attribute values and class such as [attribute_1, ⋯, attribute_n, class_i]. The learning algorithm attempts to induce from this training dataset some generalized concepts, i.e. rules that can be used to classify the remaining data. A classification scheme must be developed at this stage. The attributes to be used in learning and classification must also be determined.

The C5.0 learning algorithm first generates decision trees from the training data. These decision trees are then transformed into production rules. 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 recursive "divide and conquer" strategy is used by C5.0 to generate a decision tree from a set of training data (Quinlan, 1993). The training data set S is divided into subsets S1, ⋯, Sn according to a1, ⋯, an, which are the possible values of a single attribute A. this generates a decision tree with A being the root and S1, ⋯, Sn corresponding to subtrees T1, ⋯,

Tn. 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. If the stop condition for the procedure is satisfied, resulting in a final decision tree. The goal is to build a decision tree as small as possible. This ensures that the decision making by the tree is efficient and effective. The goal is achieved by selecting the most informative attribute at each node so that it has the power to divide the dataset corresponding to the node into pure subset as possible. Although the decision tree is an important form of knowledge representation, it is rarely used directly as knowledge base in expert systems because it is often too complex to be understood, especially when it is large. A decision tree is also difficult to maintain and update. Therefore, it is desirable to transform a decision tree to another type of knowledge representation adopted commonly in expert systems, such as production rules. Each path from the root to a leaf in a decision tree can be translated to a production rule as following:

(Attribute 1 = A, B ···), (Attribute 2 > 240),

··· → (class = forest)

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. C5.0 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 C5.0 is ordering the sets of rules for the classes according to minimized false positive errors (Quinlan,

1993). Some objects in the data to be classified may satisfy no rules. This problem can be solved by defining a default rule that assigns a default class to such objects.

4. Case study: land cover classification using inductive machine learning

1) Study area and data

The study area is "Goyang" quadrangle, No. 376082 in the standard topographic map series of South Korea at 1:25,000 scale(Figue 2). The quadrangle covers Deokyang-gu of Goyang-city and Eunpyeong-gu of Seoul-city area that are mixed with urban and rural land covers. Research data consist of the satellite image for land cover classification and GIS data layers for inductive machine learning. The remotely sensed image is leaves-on Landsat ETM+ data taken on September 4th, 2000. All spectral bands except thermal band are used for unsupervised classification. Digital elevation model (DEM) data, demographic data, hydrographic data, and transportation network data are used inductive learning.

2) Implementation

Knowledge base construction using inductive machine learning is a process to generalize rules from examples. Inductive tools use an algorithm that constructs a simple decision tree from a number of initial examples entered by the developer. This process resembles training process of supervised classification methods except attributes from various GIS data layers are used in addition to digital number (DN) values of spectral bands from a satellite image. The decision rules generated by the learning process are directly applied for land cover classification
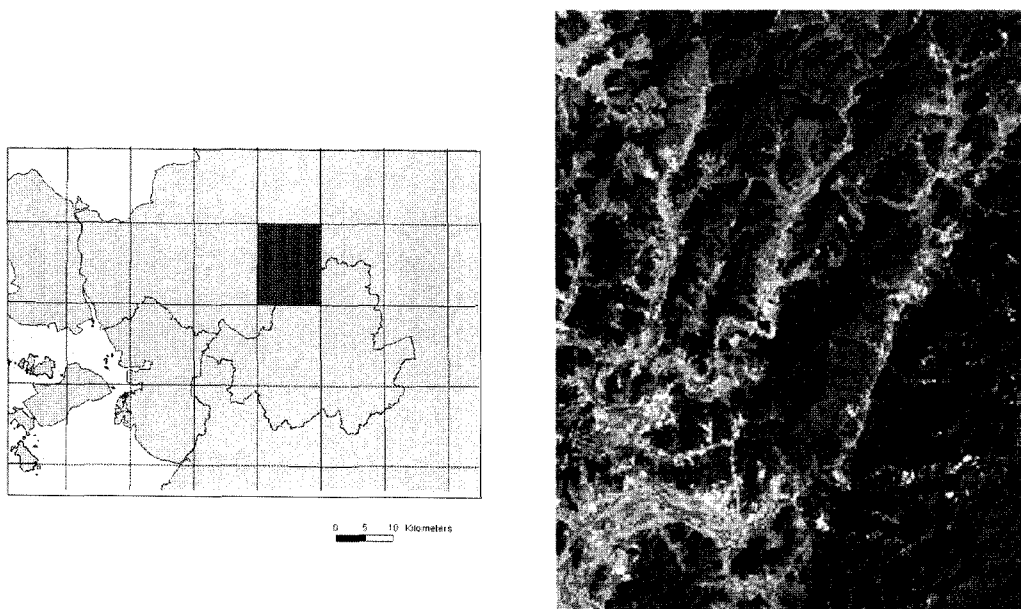
Figure 2. Study area and Data

(Left: Quadrangle No. 376082, Right: LANDSAT ETM+ Image false color composite display)
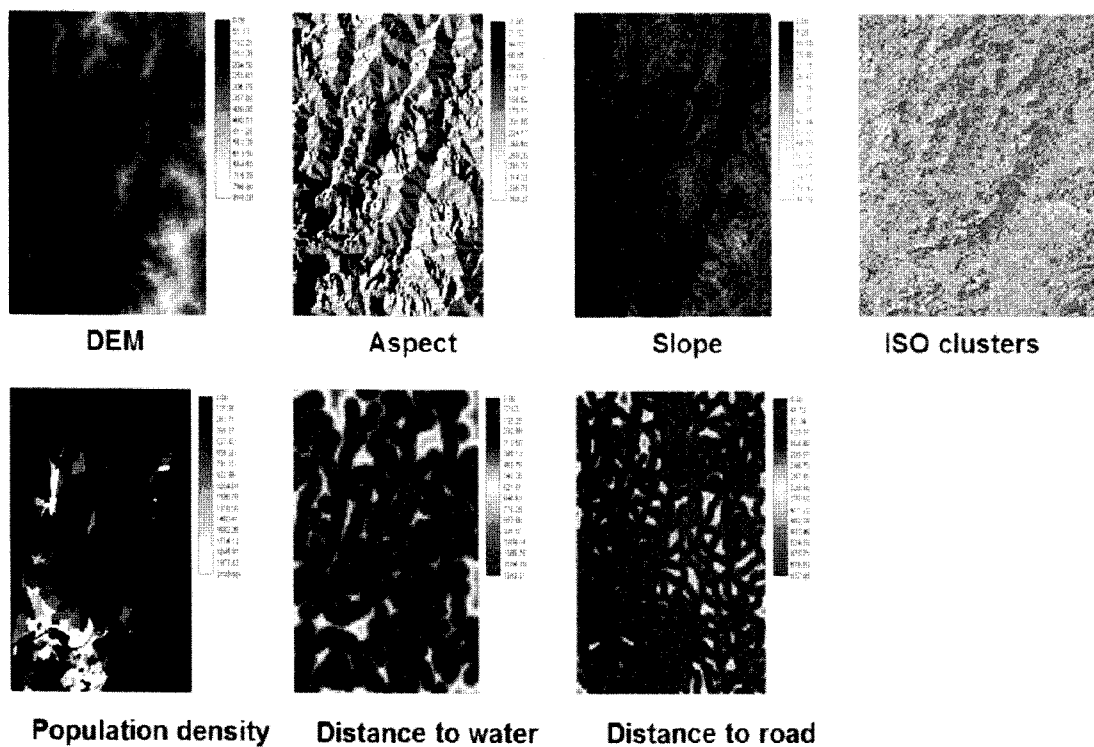

DEM   **Aspect**   **Slope**   **ISO clusters**

Population density   **Distance to water**   **Distance to road**

Figure 3. Variables for inductive learning

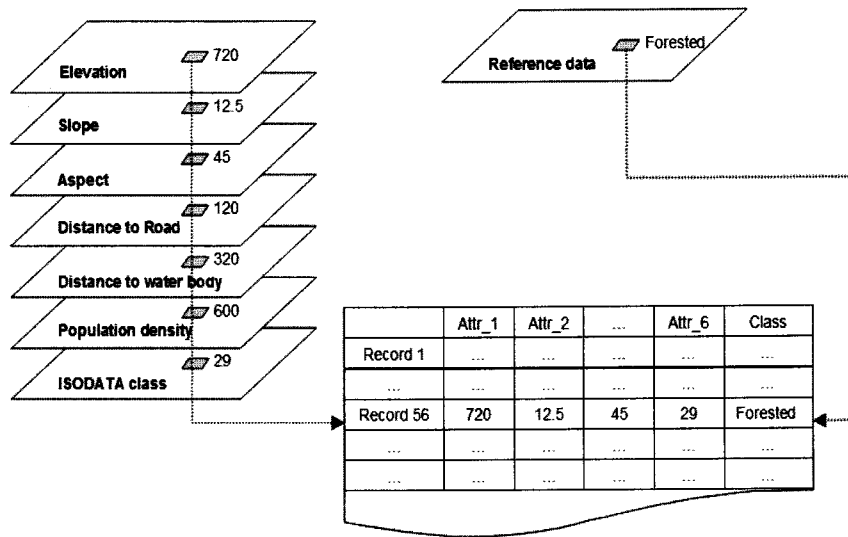|	Attr_1	Attr_2	...	Attr_6	Class
Record 1	...	...	...	...	...
...	...	...	...	...	...
Record 56	720	12.5	45	29	Forested
...	...	...	...	...	...
...	...	...	...	...	...

Figure 4. Structure of training dataset

by a series of simple "IF - THEN" operations. The process of image classification method we propose consists of five steps as following.

(1) RS image subset by unsupervised (ISODATA) classification

The Landsat ETM+ image was first pre-classified into 30 spectrally homogeneous classes using ISODATA clustering algorithm provided by ERDAS Imagine 9.3 before they were used for machine learning. Although this was not required by learning algorithm, it reduced the dimension of the spectral data from six to one.

(2) GIS data layer preparation

Aspect and slope data are generated from the DEM. Raster layers representing the distance to transportation and the distance to water bodies are generated by Euclidean distance method. Population density map by census enumeration district of the area is generated using the 2000 Population and Housing Census. All GIS data layers are converted to raster layer with 30 meter resolution to match with Landsat ETM+ image

resolution. The spectral class layer was then integrated with the other six GIS data layers to form the layer stack illustrated in Figure 3.

(3) Training data preparation

Reference land cover data for inductive machine learning were acquired from the Ministry of Environment of S. Korea. Fifty training points are randomly selected from the land cover map for each of the seven land cover classes (Forest, urban, water, agriculture, rangeland, barren land, and wetland) in Level I classification scheme defined by the Ministry of Environment of S. Korea (Refer to Environment Geographic Information System, http://egis.me.go.kr, for more information). Attribute values of each sample point are acquired by a custom Visual Basic code, then recorded as a data matrix as illustrated by Figure 4. Data matrix assembled with all input datasets and reference data is used as input for machine learning algorithm to construct the knowledge base for land cover classification.

```
DEM > 109:
:...ISO clusters in 1,2,3,4,5,6,7,9,11,13,14,15,16,17,18,19,20,21,22,23,25,27,
:   :                  28,29,30: 1 (50)
:   ISO clusters in 8,10,12,24,26: 6 (50)
DEM <= 109:
:...ISO clusters in 1,3,4,5,6,14,15,19,25,27,29: 2 (0)
    ISO clusters in 11,16,18,22:
    :...ISO clusters in 11,16:
    :   :...DEM <= 50: 7 (3)
    :   :   DEM > 50: 5 (43/1)
    :   ISO clusters in 18,22:
    :   :...DEM > 72: 4 (48/2)
    :       DEM <= 72:
    :       :...Pop density <= 115.811: 4 (3)
    :           Pop density > 115.811:
    :           :...Pop density <= 416.047: 7 (20/1)
    :               Pop density > 416.047: 5 (4)
    ISO clusters in 2,7,8,9,10,12,13,17,20,21,23,24,26,28,30:
    :...DEM <= 33: 2 (27)
        DEM > 33:
        :...Pop density > 786.136: 2 (20/1)
            Pop density <= 786.136:
            :...ISO clusters in 8,9,24: 3 (0)
                ISO clusters in 12,17,20,26: 7 (25/4)
                ISO clusters in 2,7,10,13,21,23,28,30:
                :...Distance to water <= 84.85281: 3 (52/5)
                    Distance to water > 84.85281:
                    :...Pop density <= 330.851: 2 (3)
                        Pop density > 330.851: 7 (2)
```

Figure 5. Classification decision tree generated from ISODATA clusters and GIS data

```
Rule 1: (50, lift 6.9)
        ISO clusters in 1, 2, 3, 4, 5, 6, 15, 16, 19, 27
        DEM > 109        ->  Forest  [0.981]
Rule 2: (27, lift 6.8)
        DEM <= 33        ->  Urban  [0.966]
Rule 3: (27/1, lift 6.5)
        ISO clusters in 8, 9, 10, 12, 20, 21, 24, 28, 30
        Pop density > 786.136        ->  Urban  [0.931]
Rule 4: (8, lift 6.3)
        ISO clusters in 10, 21, 28
        DEM <= 109, Pop density <= 330.851
        Distance to water > 84.85281        ->  Urban  [0.900]
Rule 5: (52/5, lift 6.2)
        ISO clusters in 2, 7, 13, 21, 23, 30
        DEM > 33,  DEM <= 109, Pop density <= 786.136
        Distance to water <= 84.85281        ->  Water  [0.889]
Rule 6: (15, lift 6.6)
        ISO clusters in 18, 22
        DEM <= 109, Pop density <= 115.811  ->  Agriculture [0.941]
Rule 7: (48/2, lift 6.6)
        ISO clusters in 18, 22
        DEM > 72        ->  Agriculture  [0.940]
Rule 8: (43/1, lift 6.7)
        ISO clusters in 11, 16
        DEM > 50, DEM <= 109        ->  Grassland  [0.956]
Rule 9: (18/2, lift 5.9)
        ISO clusters in 11, 16, 22
        DEM <= 72, Pop density > 416.047  ->  Grassland  [0.850]
Rule 10: (50, lift 6.9)
        ISO clusters in 8, 10, 12, 24, 26
        DEM > 109        ->  Barren land  [0.981]
Rule 11: (20/1, lift 6.4)
        ISO clusters in 18, 22
        DEM <= 72, Pop density > 115.811, Pop density <= 416.047
        ->  Wetland  [0.909]
Rule 12: (8, lift 6.3)
        ISO clusters in 11, 18, 22
        DEM <= 50        ->  Wetland  [0.900]
Rule 13: (25/4, lift 5.7)
        ISO clusters in 12, 17, 20, 26
        DEM > 33, DEM <= 109, Pop density <= 786.136
        ->  Wetland  [0.815]
Rule 14: (42/22, lift 3.3)
        ISO clusters in 7, 10, 12, 13, 17, 20, 21, 23, 26, 30
        Pop density > 330.851, Pop density <= 786.136
        ->  Wetland  [0.477]
```

Figure 6. Production rules generated from ISODATA clusters and GIS data

Hwahwan Kim · Cha Yong Ku

(4) Knowledge construction

The machine learning was implemented by See5 program which is developed using C5.0 algorithm by RuleQuest Research Ltd. (http://www.rulequest.com/). The input data of the learning system is a text file with each line representing a training object. The resultant decision trees and production rules are written to a file as shown in Figure 5 and 6. These decision tree and production rules serve as the knowledge base for the expert system image classification.

(5) Image classification

Applying theses production rules to image classification is straightforward. We developed a simple image classifier using Visual Basic, in which land cover class of each pixel is determined by a series of "IF-THEN" statements according to the production rules from figure 6. Figure 7 shows the result of the proposed method compared to a land cover classification based on the maximum likelihood classifier, one

of the most widely used supervised classification algorithms (Jensen, 2005).

3) Result and discussion

Table 1 shows the result of machine learning with 350 sample data points. Overall accuracy of sample training with C5.0 algorithm is 96 per cent, which means the resultant production rules predicted 96% of sample points correctly.[1] According to the contingency table and its producer's accuracy, several land cover classes are identified to be hard to discriminate from each other. Wetland, water, rangeland, and agriculture classes turned out to be commonly misclassified even though other information from various GIS layers are integrated in the training process. One common example would be the rice field of the agriculture class, which has very similar spectral signature and location close to water and wetland class so that easily confused with those classes.
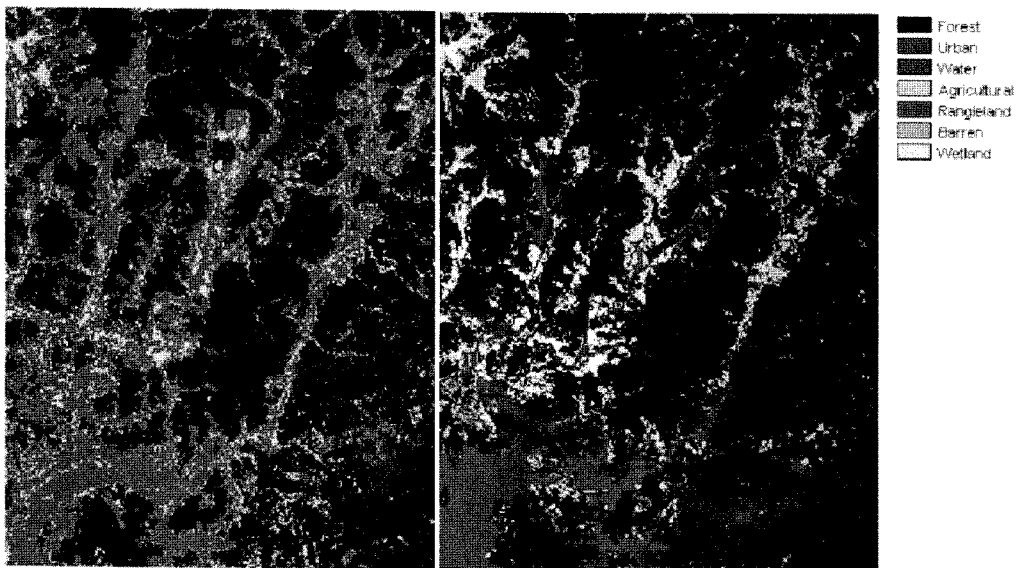

Figure 7. Classification result comparison (left: Maximum likelihood classifier, right: Inductive machine learning method)

Table 1. Evaluation on training data (350 cases)

		Classified as							Producer's Accuracy
		Forest	Urban	Water	Agriculture	Rangeland	Barren	Wetland	
	Forest	50							100%
	Urban		49	1					98%
Reference	Water		1	47				2	94%
	Agriculture				49			1	98%
	Rangeland				2	46		2	92%
	Barren land						50		100%
	Wetland			4		1		45	90%
	Overall								96%

C5.0 algorithm also provides information about how each of input attributes contributes to the determination of production rules as 'Attribute usage' measure as given in table 2. Among the seven input data layers (one pre-classified ISODATA cluster and six GIS data layers), only four layers significantly contribute to the generation of production rules according to the training results. Two bi-products of elevation data (aspect and slope) and distance to road network turned out to be irrelevant to land cover classification according to the training result.

Table 2. Attribute usage

Input layer	Usage
ISO clusters	96%
DEM	94%
Population density	48%
Distance to water	17%

Determining the usage of each attribute in the training process, inductive machine learning algorithm provides the capability to assess what kind of additional GIS data layers can be effectively utilized to improve the performance of land cover classification. This capability is one of the most promising advantages the inductive machine learning algorithm has for image classification research.

There are several land cover types which are easy to discern each other in traditional classification using spectral characteristics only. Among the thirty ISODATA clusters pre-classified with special signatures, about a half of those are classified as a certain land cover class straightforwardly. For example, ISODATA clusters of 9, 14, 25, 28, and 29 are classified as 'Urban' without exception according to the production rules of Figure 6. Also ISODATA clusters of 1, 3, 4, 5, 6, 15, 19, and 27 are classified as 'Forested'. This means that the spectral characteristics of these two land cover classes are relatively unique compared to other classes.

Otherwise, land cover classes of the remaining ISODATA clusters are determined by the information from the additional GIS data layers in addition to spectral signature. For these cases, spectral characteristics summarized in the form of ISODATA cluster are not enough to discern a certain land cover class from other classes. For example, pixels of ISODATA clusters of 8, 10, 12, and 24 have spectral characteristics that can be classified as both of 'Barren land' and 'Urban', according to the training data. Also, ISODATA clusters of 7, 21, 23, and 30 have very similar spectral characteristics common with 'Water' and 'Wetland' class concurrently. Therefore, spectral data alone were not capable of distinguishing

these classes from each other. The machine learning approach obtained significant improvement for these classes. From the decision tree (Figure 5) and production rules (Figure 6) generated from this approach, it is obvious that the GIS data layers played an important role in the improvements.

For example, DEM, distance to water body and population density were used to distinguish 'Urban' land cover from 'Barren land' and 'Water' classes the following rules:

Rule 4: ISO clusters in 10, 21, 28

DEM <= 109, Pop density <= 330.851, Distance to water > 84.85281 -> Urban

If a pixel's ISODATA cluster value is among the group of Rule 4, but its GIS attributes do not satisfy the conditions of Rule 4, this pixel is classified as to 'Barren land' or 'Water' class according to another rule sets.

One advantage of the land cover classification based on machine learning algorithm is its semi object-based approach. Compared to the result of Maximum likelihood classification method (Figure 7 left), our method produces clearly delineated thus less granulated land cover map. Our method pre-classifies the original Landsat ETM+ image into spectral clusters before the training process. By this process, pixels with similar spectral characteristics are grouped into a number of clusters. Although pixels in the same spectral cluster can be classified as different land cover class in the final classification, the membership to a spectral cluster of each pixel works as one of major factor in determining the land cover class of the pixel.

Another advantage of the proposed method is in its potential to improvement. On the contrary to the neural network method for land cover classification, machine learning algorithm provides intuitive production rules researchers can easily understand and improve. Any GIS data layer can be included in the proposed method

based on the test to see how those attributes contribute. Integration of various GIS data and domain knowledge with the automated classification can be regarded as the most important advantage this method can provide for better land cover image classification.

5. Conclusion

Integration of GIS data and human expert knowledge into digital image processing has long been acknowledged as a necessity to improve remote sensing image analysis. One of the most encouraging solutions to the problem is mining knowledge from spatial datasets and utilizing the knowledge in image interpretation for spatial data updating. The implementation of inductive learning in spatial databases and the combination with traditional classification methods are theoretically and practically valuable. In this research, we proposed inductive machine learning algorithm for GIS data integration and rule-based classification method for land cover classification. 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 knowledge engineers to code the knowledge. In order to verify the feasibility of the inductive learning based knowledge construction, a land cover classification experiment was performed with a Landsat ETM+ multispectral image and GIS data layers including elevation, aspect, slope, distance to water bodies, distance to road network, and population density. Decision trees and production rules for land cover classification were generated by C5.0 inductive machine learning algorithm with 350

stratified random point samples, and used for land cover classification integrated with unsupervised ISODATA classification. Result shows that GIS data layers such as elevation, distance to water bodies and population density can be effectively integrated for rule-based image classification. This study demonstrated the utility of GIS data to improve land cover classification of remotely sensed imagery. GIS data usually do not meet the Gaussian distribution assumption, conventional classification method such as maximum likelihood method may not be appropriate for land cover classification based on GIS data integration. On the other hand, the expert system approach proved to be an effective way to incorporate GIS data because it does not have such a data distribution requirements. The research also demonstrated some other advantages of the machine learning approach. It is easy to understand, and the resultant knowledge base could be used in other applications. Such spatial knowledge would be useful in many geographic applications such as spatial analysis and modeling.

Proposed method demonstrates how various GIS data layers can be integrated with remotely sensed imagery in a framework of knowledge base construction to improve land cover classification.

Note

1) It's worth nothing that the contingency matrix reported here is not for the classified land cover map but for the machine learning training. User's accuracy is therefore not reported explicitly. This paper focused on the implementation of machine learning algorithm based knowledge base construction for expert system land cover classification. Appropriate accuracy assessment of the classified land cover map is going to be reported by future research although it was not implemented in this research due to limited data availability.

Reference

Ball, G. H. and Hall, D. J., 1965, *A Novel Method of Data Analysis and Pattern Classification*, Menlo Park, Stanford Research Institute.

Eklund, P. W., Kirkby, S. D., and Salim, A., 1998, Data mining and soil salinity analysis, *International Journal of Geographical Information Science*, 12(3), 247-268.

Enslin, W. R., Tonand, J., and Jain, A., 1987, Land cover change detection using a GIS-guided feature-based classification of Landsat Thematic Mapper data, *Proc. ASPRS*, 6, 108-120.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., 1996, *Advances in knowledge Discovery and Data Mining*, AAAI/MIT Press, Menlo Park.

Gahegan, M., 2000, On the application of inductive machine learning tools to geographical analysis, *Geographical Analysis*, 32, 113-139.

Gahegan, M., 2003, Is inductive machine learning just another wild goose (or might it lay the golden egg)?, *International Journal of Geographic Information Science*, 17(1), 69-92.

Huang, X. and Jensen, J. R., 1997, A Machine Learning approach to automated knowledge-base building for remote sensing image analysis with GIS data, *Photogrammetric Engineering and Remote Sensing*, 63(10), 1185-1194.

Jensen, J. R., 2005, *Introductory digital image processing: A Remote Sensing Perspective*, Upper Saddle River, Prentice Hall, New Jersey.

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

Li, D., Di, K., and Li, D., 2000, Land use classification of remote sensing image with GIS data based on spatial data mining techniques, *International Archives of Photogrammetry and Remote Sensing*, 33, Part B3.

Malebra, D., Esposito, F., Lanza, A., and Lisi, F. A., 2001, Machine learning for information extraction from

topographic maps. In Miller, H. J. and Han, J. (Eds.), *Geographic Data Mining and Knowledge Discovery* (pp. 291-314). Taylor and Francis, New York.

Michalski, R. S. and Chilausky, R. L., 1980, Learning by being told and learning from examples: an experimental comparison of the two methods of knowledge acquisition in the context of developing and expert system for soybean disease diagnosis, *Policy Analysis and Information Systems*, 4(2), 1980.

Mitchell, T. M., 1997, *Machine Learning*, McGraw-Hill, New York.

Qi, F. and Zhu, A. 2003, Knowledge discovery from soil maps using inductive learning, *International Journal of Geographical Information Science*, 17(8), 771-795.

Quinlan, J. R., 1986, Induction of decision trees, *Machine Learning*, 1, 81-106.

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

Quinlan, J. R., 2001, See5: *An Informal Tutorial*, Accessed at URL: http://www.rulequest.com

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

Correpondence: Cha Yong Ku, Department of Geography, Sangmyung University, Hongji-dong, Jongro-gu, 110-743, Seoul, Korea (e-mail: koostar@smu.ac.kr, phone: 02-2287-5043)

교신: 구자용, 서울시 종로구 홍지동 상명대학교 사회과학부 지리학전공, 110-743 (이메일: koostar@smu.ac.kr, 전화: 02-2287-5043)