한국지리정보학회지 7권 1호 : 9~19(2004)

의사결정트리를 이용한 개별 공시지가 비교표준지의 자동 선정

김종윤[1※] · 박수홍[2]

An Automatic Method for Selecting Comparative Standard Land Parcels in Land Price Appraisal Using a Decision Tree

Jong-Yoon KIM[1※] · Soo-Hong PARK[2]

요 약

개별 공시지가 산정에 있어 비교 표준지의 선정은 가장 중요한 작업으로서, 최대한 객관적이고 합리적으로 이루어져야 한다. 그러나 현재 비교표준지를 선정하는 작업은 담당 공무원의 수작업에 의해 이루어지기 때문에 효율성이나 객관성을 보장하기가 어렵다. 본 연구에서는 현행 비교표준지 선정방식을 분석하여 문제를 정의하고 비교표준지 선정 업무의 자동화에 적용가능한 기계학습 알고리즘으로 의사결정트리를 선정하고 비교표준지를 선정하여 규칙을 주제지향적인 데이터베이스를 기반으로 학습하였다. 이렇게 학습된 규칙을 이용하여 비교표준지를 선정하고 그 결과를 평가 분석하여 새로운 비교표준지 선정 방법을 제안하였다.

주요어 : 개별공시지가, 비교표준지, 의사결정트리

ABSTRACT

The selection of comparative standard parcels should be objective and reasonable, which is an important task in the individual land price appraisal procedure. However, the current procedure is mainly done manually by government officials. Therefore, the efficiency and objectiveness of this selection procedure is not guaranteed and questionable. In this study, we first defined the problem by analyzing the current comparative standard land parcel selection method. In addition, we devised a decision tree-based method using a machine learning algorithm that is considered to be efficient and objective compared to the current selection procedure. Finally the proposed method is then applied to the study area for evaluating the appropriateness and accuracy.

KEYWORDS : Individual Public Land Price, Comparative Standard Land Parcel, Decision Tree

2003년 11월 17일 접수 Recieved on November 17, 2003 / 2004년 3월 19일 심사완료 Accepted on March 19, 2004

1 (주)지노시스템, G-inno Systems
2 인하대학교 지리정보공학과 Department of Geoinformatic Engineering, Inha University
※ 연락저자 E-mail: jykim@g-inno.com

서 론

1. 연구배경과 목적

토지의 가격 결정 과정은 일정한 기준에 의해 토지의 가치를 평가하는 작업으로서 도시 지역의 경우 도시 공간구조 분석을 위한 중요한 지표로 활용되고(채미옥, 1997), 지가의 객관적이고 정확한 평가를 통해 토지시장의 안정화를 꾀할 수 있으며, 토지관련 국세 및 지방세 부과기준으로 활용됨은 물론 개발 부담금 등 각종 부담금의 부과기준으로 활용된다(홍길순, 1998). 그러나 2,700만 필지에 대해 약 두 달 동안 지가 담당 공무원들에 의해 이루어지는 개별공시지가산정 작업은 방대한 예산과 시간이 소요되기 때문에 자동화에 대한 요구가 있어왔다.

이러한 문제를 해결하기 위하여 건설교통부에서는 개별지가 자동 산정 프로그램인 ALPAS (Automatic Land Price Appraisal System)를 개발하여 개별지가를 자동으로 산정하고 있다. 하지만 ALPAS는 조사된 토지특성을 지가산정방식에 따라 계산하는 방식으로 단순히 산정과정을 전산화한 것에 불과하며, 토지특성 추출에 관한 기능과 비교표준지에 대한 취사선택 기능이 없는 등 정확도나 효율성 면에서 많은 문제점을 드러내고 있다(채미옥과 권태형, 1997; 문태헌, 2000)

비교표준지를 이용하여 개별공시지가를 산정하는 우리나라 제도 하에서 가장 중요한 문제는 개별필지 주변의 표준지 중에서 어떤 표준지를 선택·이용하여 지가를 산정해야 하는가이다. 현재 비교표준지의 선정은 지가담당 공무원(실무자)들에 의해 이루어지고 있다. 그러나 담당 공무원들의 전문성 부족, 주관성의 개입 등으로 인하여 비교표준지 선정 업무가 정확도와 객관성 측면에서 많은 문제점을 드러내고 있다(하철영과 박정호, 2000). 이러한 문제점들은 지가관련 행정업무에 대한 민원을 증가시켜 공시지가 산정에 대한 신뢰성을 저하시키고 있다. 이에 보다 객관적이고 합리적인 기준에 의해 비교표준지를 정확하게 선정할 수 있는 방법론이 요구되고 있다.

본 연구에서는 보다 효율적이고 객관적인 방법으로 비교표준지를 선정하고자 기계학습 알고리즘의 하나인 의사결정트리를 사용하여 개별공시지가를 산정할 때 비교표준지를 선정하는 규칙 또는 패턴을 추출한 후, 이렇게 학습된 규칙을 이용하여 비교표준지를 자동 선정한 후 현행 비교표준지와 비교 분석을 통하여 본 연구에서 제시하는 방법론의 정확성과 타당성을 평가하였다.

2. 연구내용과 방법

본 연구에서는 현행 개별공시지가 산정 절차와 비교표준지 선정 방법을 분석을 통하여 문제를 정의하고, 비교표준지 선정 업무의 자동화에 적용 가능한 기계학습 알고리즘을 탐색하였다. 적합한 기계학습 알고리즘을 선정한 후 비교표준지 선정규칙의 학습을 위한 데이터웨어하우스를 구축하고 이에 선정된 기계학습 알고리즘을 적용하여 비교표준지 선정 규칙을 도출하였다. 이 규칙을 실제 공시지가 비교표준지 선정에 적용하기 위하여 ArcObjects를 사용하여 프로토타입 형태의 비교표준지 자동선정 시스템을 구현하였다.

연구지역은 서울특별시 강남구의 삼성동과 대치동을 선정하였으며, 용도지역은 각 동의 제1종 전용주거지역과 일반상업지역을 선정하였다. 비교표준지 선정의 경우에는 정확도와 관련하여 절대적인 기준이 없으므로, 현행 비교표준지와의 비교 분석을 통하여 자동선정 방법론의 정확도와 타당성을 평가하였다.

비교표준지 선정 방법과 의사결정트리

1. 현행 비교표준지 선정방법 및 문제점

비교표준지란 개별 공시지가를 산정하고자

하는 필지주변의 여러 표준지 중에서 직접 비교의 기준이 되는 표준지(행정구역 경계지역에서는 인접지역 비교표준지 선정 가능)로서 산정 대상 필지와 토지특성 비교를 통하여 비준율을 적용하게 되는 표준지를 말한다. 선정 방법은 산정 대상 토지가 일반 토지인 경우에는 조사대상 토지와 동일 용도지역(개발제한구역 포함)안에 있는 유사가격권의 표준지 중에서 조사대상 토지와 토지이용상황이 같거나 도로접면이 유사한 표준지를 선정하며, 동일한 용도지역내 토지이용상황이 같은 유사가격권의 표준지가 2개 이상일 경우에는 주용도 내의 세항이 같은 표준지 1개를 선택한다. 동일한 용도지역내 토지이용상황(주용도)이 같은 유사가격권의 표준지가 없는 경우에는 주용도가 다르더라도 조사대상필지 인근의 토지이용상황을 감안하여 유사가격권의 표준지를 선정한다(건설교통부, 2001).

이렇게 비교방식에 의한 지가산정에서는 산정의 기준이 되는 비교 표준지를 적절하게 선정하는 것이 무엇보다 중요하며, 비교표준지 선정시 가장 유의할 사항은 선정자의 임의성을 배제하고 비교 표준지 선정기준에 의거하여 합리적인 선정이 이루어지도록 해야 한다. 그러나, 현재의 비교표준지 선정은 지가담당 공무원들에 의해 수작업으로 이루어지고 있어 문제점이 발생하고 있는데, 구체적으로 담당 공무원의 빈번한 인사 교체로 인해 지가조사 경험과 지식이 축척되지 않음으로 말미암아 오류가 발생하고 있다. 또한 유사가격권내 또는 동일 용도지역의 토지이용상황이 같은 표준지가 여러 개 있을 경우, 조사 담당공무원들이 정확하게 토지특성과 거리 등을 고려하여 최적의 비교표준지를 선정하기란 현실적으로 한계성이 있다(홍길순, 1998; 최양춘, 2001). 이러한 문제점을 개선하기 위한 방안으로 비교표준지 4대 선정원칙을 구체화하고 토지이용상황별 요건을 세부적으로 제시(정수연, 2002)하고 있지만 그 기준이 여전히 모호하다. 이러한 문제점에 기인하여 개별공시지가를 검증하는 과정에서 오류가 발견되는 필지 중 비교표준지의 선정 착오가 대략 40%(국토개발연구원, 1997)로 나타나고 있어 정확하고 객관적으로 비교표준지를 선정할 수 있는 자동화된 방법론이 요구되고 있다.

2. 기계학습 알고리즘의 종류

기계학습 알고리즘은 데이터마이닝 분야에서 전통적인 통계기법과 함께 주로 사용되는 방법으로서 이 절에서는 비교표준지 선정에 활용할 알고리즘을 선정하기 위해 대표적인 기계학습 알고리즘들, 즉 1R, 베이지안망, 신경망, 의사결정트리에 대해 고찰하고자 한다. 1R (1-Rule) 학습시스템은 극히 단순한 방법으로도 상당한 수준의 학습결과를 얻을 수 있다. 이것이 취하고 있는 기본적 방법은 주어진 예들을 분류하는 데에는 많은 속성이 필요 없고 단 하나의 속성만을 시험하는 것이다(Holte, 1993). 1R은 기존의 학습 알고리즘들에 대한 대안으로 개발된 것은 아니며, 최소한 이러한 방법보다는 나아야 한다는 기준을 제시해 주는 장치로서 의미를 지니고 있다(류광렬, 2002).

일반적으로 베이지안망은 전산학적인 측면에서 여러 관심 대상들 간의 관련성에 대해 확률적으로 표현함으로써 그 구조 및 의존성을 잘 표현할 뿐만 아니라 수학적인 면에서도 견고성이 입증된 통계적인 도구이다(Han, 2001). 베이즈 이론은 사전적 가설들과, 관측된 데이터가 주어질 때, 관심대상이 되는 다양한 데이터의 확률과 사전 확률에 의거하는 가설들의 사후 확률을 계산할 수 있는 방법을 제시한다. 베이지안망은 이 베이즈 이론을 기초로 변수간의 확률관계를 비교적 축약된 형태로 나타내는 그래프 모델이다. 베이지안망은 기본적으로 변수들의 결합 확률분포를 나타내기 때문에 모델의 예측과 설명이 가능하며 각 변수들에 대한 의존성까지도 함께 확률로서 표현된다(Heckerman, 1996).

신경망은 인간두뇌의 신경망 세포를 모방한 개념으로 마디(node)와 고리(link)로 구성된 망 구조를 모형화하고, 과거에 수집된 데이터로부터 반복적인 학습 과정을 거쳐 데이터에 내재되어 있는 패턴을 찾아내는 모델링 기법으로 신용평가, 카드 도용패턴 분석, 수요 및 판매 예측, 고객세분화 등 여러 가지 목적으로 다양한 분야에 폭 넓게 적용되고 있다(문태헌, 2000; 장남식 등, 2002). 인공신경망은 신경생리학 분야에서 두뇌의 활동을 이해하고자 하는 목적하에 신경의 작업을 설명하려는 시도에서 출발하였으며 데이터의 증가와 그 안에 잠재되어 있는 정보를 얻는 작업이 더욱 어렵게 됨으로 인해 다양한 통계분석의 보완적 방법으로 새로운 관심을 받게 되었다. 인공신경망 분석은 방대한 데이터들로부터 중요한 패턴을 찾고, 분류하고 연관성을 살피고 군집으로 나누기 위해 필요한 정보의 획득을 보다 효과적으로 수행할 수 있게 만드는 것이다. 이를 보유하고 있는 대규모의 데이터베이스에서 표면적으로 드러나지 않고 내재되어 있는 정보들을 도출해내는 기술인 데이터분석이 이용되면서 최근에 각광을 받고 있다(안지현, 2002).

Quinlan에 의해 제시된 C4.5는 대표적인 의사결정트리 학습 기법으로, 현재 가장 널리 사용되고 있는 학습 기법들 중의 하나다. 의사결정트리는 트리의 구조에 기반한 분류/예측 모델로서 과거에 수집된 데이터의 레코드들을 분석하여 이들 사이에 존재하는 패턴, 즉 분류별 특성을 속성의 조합으로 나타내는 분류모형을 나무의 형태로 만드는 것이다. 그리고 이렇게 만들어진 분류모형은 새로운 레코드들을 분류하고 해당 부류의 값을 예측하는데 사용된다(장남식 등, 2002). 의사결정트리는 트리형태의 결과를 제시해 주므로 그 해석이 용이하며, 목표변수에 대한 두개 이상의 변수의 복합적인 영향관계를 분석할 수 있다. 또한 선형성, 정규성, 등분산 등의 가정을 필요로 하지 않는 비모수적인 방법이며 의사결정트리의 각 노드로부터 설명 가능한 규칙을 생성할 수 있다는 특징을 갖는다(Berry와 Linoff, 1997). 레코드들을 분류하기 위해서는 루트노드부터 시작하여 중간노드들을 거쳐 각 노드마다 하나씩 지정된 속성을 검사하면 된다. 각 노드에서 속성을 검사한 결과값에 따라 지정된 가지를 따라 가다가 잎 노드에 도달하면 클래스가 결정된다. 결정트리 학습의 문제는 바로 이런 결정트리를 주어진 훈련집합으로부터 유도해 내는 것이다. 일반적으로 집합 S가 n개의 서로 다른 클래스에 속하는 사례들로 구성되어 있을 때, p_i를 S 내에서 i번째 클래스가 차지하는 비율이라 한다면 S의 불순도(impurity)의 척도인 엔트로피(entropy)는 다음과 같이 정의된다.

$$E(S) \equiv \sum_{i=1}^{n} - p_i \log_2 p_i$$

이 엔트로피 값을 이용하면 분류를 위한 적합한 속성을 쉽게 찾을 수 있다. 속성의 값에 따라 사례들을 나누었을 때 불순도, 즉 엔트로피의 값이 크게 낮아지는 속성을 찾으면 되는 것이다. 정보이론에서는 엔트로피의 감소량을 정보이득(information gain)이라 부르고 있다(Han과 Kamber, 2001). 예의 집합 S를 속성 A의 값에 따라 n개로 나누었을 때 얻게 되는 정보이득은 다음과 같이 계산된다.

$$Gain(S, A) \equiv E(S) - \sum_{k=1}^{n} \frac{|S_k|}{|S|} E(S_k)$$

이 식에서 첫 항 E(S)는 나누기 전의 S의 엔트로피이고 둘째 항은 나눈 후 생기는 각 부분집합의 엔트로피를 평균한 것이다. 단, 부분집합의 크기가 서로 다르므로 단순 산술평균이 아니라 각 부분집합의 엔트로피 값 $E(S_k)$에 그 부분집합이 S에서 차지하는 비율 $|S_k|/|S|$를 가중치로 곱하여 합산함으로써 구한 것이다. Gain(S, A)는 분류를 위해 속성 A가 제공하는 정보량이라고도 해석할 수 있다. 따라서 정보

이득이 가장 큰 속성을 선택하는 것이다(Han과 Kamber, 2001).

3. 비교표준지 선정규칙 학습을 위한 알고리즘 선정

비교표준지 선정을 위한 규칙의 학습은 여러 가지 조건(속성)을 고려하여, 그 결과를 이진(binary)형태로 도출하는 개념학습(concept learning)이다. 이와 같은 경우에는 '비교 표준지로 선정'이 개념에 해당하며 학습의 결과로 비교표준지로 선정되는 조건이 유도된다. 이는 '선정' 또는 '선정되지 않음' 이라는 두 개의 클래스를 갖는 분류(classification)의 문제이다.

앞 절에서 살펴보았듯이 1R은 학습을 위해 단 하나의 속성만 사용하는 것으로써 최소한 이보다는 나아야 한다는 기준을 제시하는 정도로써 사용되고 있을 뿐이다. 베이지안 망은 확률적 추론에 의한 방식으로 학습을 하는데, 조건(속성들의 값)에 의해서 표적개념이 발생하는 확률만을 표현하기 때문에 규칙을 학습하고 그 결과를 표현하기 위한 알고리즘으로는 적합하지 않으며 대용량 데이터에서의 베이지안 망 구조 학습의 문제를 갖고 있다. 이는 변수간의 상관관계를 나타내는 연결선(edge)인 망을 데이터로부터 직접 학습하는 것으로 초기의 베이지안 망 사용자들은 대부분 사전지식을 이용해서 베이지안망의 구조를 고정해놓고 이를 기반으로 파라미터를 데이터로부터 학습한 후 예측이나 추론등의 과제에 베이지안망을 사용해왔다(하선영, 2001). 이는 분석가에게 데이터에 대한 충분한 사전지식을 요하는 것으로 사용자의 개입이 필요하기 때문에 연구의 목적인 비교표준지 자동선정에는 적합하지 않다. 또한 베이즈 이론은 변수간에 조건부 독립성을 가정하고 있는데 연구에서 사용하게 될 속성인 '지가', '도로접면조건', '토지이용' 등은 서로 종속성이 존재하기 때문에 베이지만 망의 사용은 적절하지 않다.

개별필지에 대한 비교표준지 선정은 필지의 가격을 평가하는데 절대적인 기준이 되기 때문에 그 근거는 명확할 필요가 있다. 그러나 인공신경망의 경우 분류의 결과만을 제공할 뿐 분류의 조건이 제공되지 않는 블랙박스적 방법이기 때문에 여러 속성들 간의 관계를 밝혀내는 데는 적합하지 않다(안지현, 2002). 따라서 인공신경망을 통한 비교표준지의 선택은 그 근거를 알 수 없기 때문에 비교표준지 선택의 업무에는 적절하지 못하다.

반면, 의사결정트리는 범주화(classification)와 예측(prediction)을 위한 강력하고도 널리 유용하게 사용되는 수단으로, 이 방법의 장점은 인공신경망과는 대조적으로 이해 가능한 규칙(rule)을 만들어 낸다는 사실에 있다. 규칙은 언어로 표현할 수 있어서 자연어 또는 SQL과 같은 database access language로 표현하여 특정 범주의 레코드를 손쉽게 추출할 수 있도록 한다(김준회, 2001). 특히 의사결정트리는 결과의 예측이 판별(분류)의 형태를 취하고 있을 때 선택하는 것이 좋다(정 현, 1999).

이러한 특성으로 인해 확률적 결과만을 보여주는 Naive Bayes 분류기나 분류결과 만을 제공하는 인공 신경망을 비교표준지 선정 업무에 적용하기에는 무리가 있다고 판단된다. 분류결과에 따른 속성의 조건, 다시 말해 토지특성 조건을 제공하며, 어느 속성이 비교표준지를 결정하는데 가장 큰 지표가 되는지를 제공하고 데이터 웨어하우스 형태로 구축된 데이터에 손쉽게 접근이 가능한 의사결정트리를 사용하는 것이 타당하다고 판단된다.

실험 및 분석

1. 실험

본 연구는 강남구 삼성동과 대치동의 '제1종 전용주거지역', '일반상업지역'을 실험지역으로 선정하였으며, 삼성동 제1종 전용주거지역 294개 필지와 일반상업지역 386개 필지, 대치동 제1종 전용주거지역 76개 필지와 일반상업

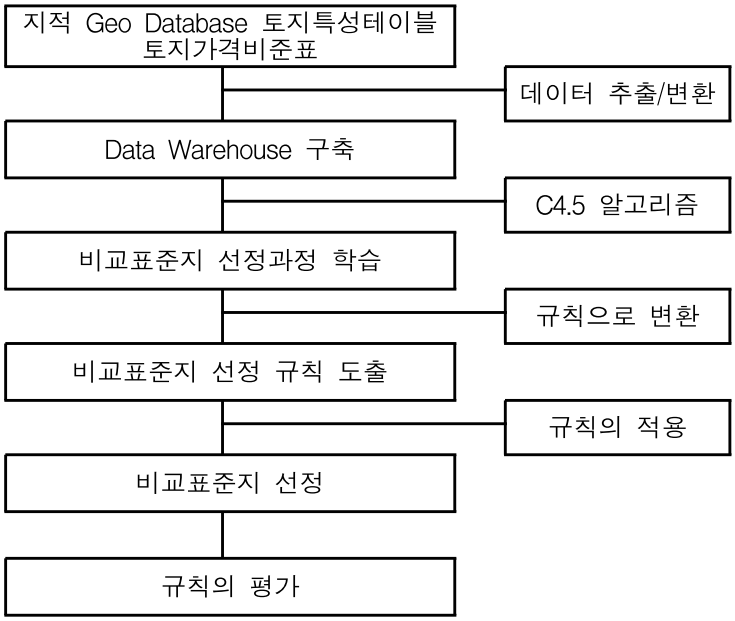


FIGURE 1. The flowchart of this study

지역 218개 일반개별필지를 대상으로 비교표준지 선정 규칙을 학습하고, 이를 적용하였다. 또한, 지식발견 프로세스를 최대한 반영하여 진행하였다. 그림 1은 전체적인 실험절차를 나타낸 것이다.

실무자가 개별필지의 비교표준지를 선택하기 위해서는 동일한 용도지역을 대전제로 조사대상 필지 주변의 비슷한 가격대, 즉 유사가격권내의 표준지들을 선정한 후 토지이용상황이 같거나 도로접면이 유사한 표준지를 선정하여 비교표준지를 선택하게 된다. 이러한 패턴 또는 규칙성을 찾아내기 위해서 본 연구에서는 의사결정트리 알고리즘인 C4.5를 사용하여 비교표준지를 선정하는 규칙을 도출하고자 한다. 이를 위해서는 각각의 개별필지와 실제 비교표준지를 포함하는 후보 비교 표준지들 간의 토지특성차이를 표현하는 형태의 데이터가 구축되어야 한다. 즉, 하나의 개별필지와 후보 비교표준지들 간의 '토지이용의 차이', '지가의 차이', '도로접면 조건의 차이', '거리차이', '실제 표준지여부'를 속성으로 사용하는 형태의 데이터가 되는 것이다. 이 데이터를 실제 표준지 선정여부를 나타내는 속성을 분류대상으로 하여 의사결정트리를 구축하면 개별필지와 후보비교표준지간의 토지특성차이에 의한 비교표준지 선정규칙을 도출할 수 있다. 개별필지와 후보비교표준지간의 토지특성차이를 나타내는 주제지향적(subject-oriented)인 데이터 웨어하우스를 구축하고 규칙의 학습/적용을 위한 시스템은 ArcObjects를 사용하여 구현하였다.

실험을 위한 원시데이터는 수치지적도와 토지특성테이블을 조인하여 개별필지의 토지특성을 포함하는 공간데이터베이스와 관계형 테이블 형태로 구축된 토지가격 비준표가 사용되었다. 토지가격 비준표란 공시된 표준지들의 토지특성을 다중회귀분석하여 추출된 토지특성별 배율을 행렬표 형태로 재구성한 표를 말한다. 이를 이용하여 위에서 언급한 '도로접면', '토지이용'에 따른 필지 간 차이를 0~1 사이의 배율로서 나타낼 수 있다.

앞에서 언급했듯이 비교표준지 선정을 위한 규칙의 학습은 여러가지 토지특성 조건을 고려

하여 그 결과를 이진형태(표준지/표준지가 아님)로 도출하는 개념학습이다. 규칙의 학습을 위해서는 원시데이터를 필요한 형태의 데이터로 변환하는 과정이 필요하며 이는 다음과 같은 단계를 거쳐 수행된다.

① 첫 번째 개별필지에 대해서 일정한 거리의 버퍼를 생성한다.

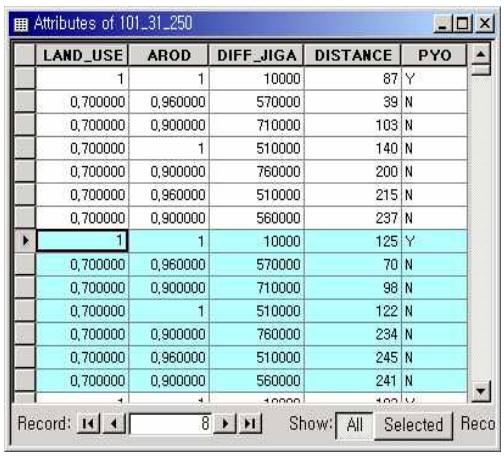
② 버퍼 내에 존재하는 표준지들 중에서 용도지역이 동일한 표준지들만 선택하여 후보 비교표준지들을 선정한다. 이 때 실제 표준지가 반드시 선택되어야 한다.

③ 개별필지와 선택된 후보 비교표준지들 각각에 대해서 토지이용 차이, 도로접면 차이, 지가의 차이, 거리 차이를 구하고, 실제 표준지 선정 여부를 판별한다. 이때, 토지이용 차이와 도로접면 차이는 해당지역의 토지가격 비준표를 사용하여 구한다.

④ ③의 결과를 관계형 데이터베이스의 테이블에 기록한다.

⑤ 모든 조사대상 필지들에 대해 ①~④의 과정을 반복한다.

위의 과정을 통하여 구축된 데이터는 그림 2와 같다.


Attributes of 101_31_250

LAND_USE	AROD	DIFF_JIGA	DISTANCE	PYO
1	1	10000	87	Y
0,700000	0,960000	570000	39	N
0,700000	0,900000	710000	103	N
0,700000	1	510000	140	N
0,700000	0,900000	760000	200	N
0,700000	0,960000	510000	215	N
0,700000	0,900000	560000	237	N
1	1	10000	125	Y
0,700000	0,960000	570000	70	N
0,700000	0,900000	710000	98	N
0,700000	1	510000	122	N
0,700000	0,900000	760000	234	N
0,700000	0,960000	510000	245	N
0,700000	0,900000	560000	241	N

Record: 8 Show: All Selected Reco

FIGURE 2. Extracted/transformed data set from original data

그림 2의 테이블에서 LAND_USE는 토지이용의 차이, AROD는 도로접면의 차이, DIFF_JIGA는 지가의 차이, DISTANCE는 개별필지와 후보 비교표준지 사이의 거리, PYO는 실제 표준지 선정 여부를 나타낸다. 그림 2에서 선택된 레코드들은 하나의 개별필지에서 선택된 후보 비교표준지들을 의미하는데, 토지이용과 도로접면이 동일하고 지가의 차이가 10,000원이며, 선정대상 개별필지와의 거리가 125m인 표준지가 비교표준지로 선정되었다는 것을 알 수 있다. 위의 데이터셋을 기반으로 비교표준지 선정 규칙을 학습하기 위해 실제 표준지 여부를 나타내는 속성을 분류대상으로 C4.5 알고리즘을 사용하여 의사결정트리를 만들고, 이를 IF~THEN 규칙의 형태로 변환한다. 학습되는 규칙은 아래 그림 3과 같은 형태를 갖는다.

의사결정트리를 통해 학습한 규칙을 적용하는 방법은 다음과 같다.

① 비교표준지 선정대상 개별필지를 선택한다.

② 개별필지로부터 일정거리 내에 있는 후보 비교표준지들을 선택한다.

③ 개별필지와 n개의 후보 비교표준지 간의 토지이용, 도로접면, 지가, 거리차이를 구한다.

④ IF~THEN 규칙의 형태로 변환된 비교표준지 선정규칙을 적용하여 ②에서 선택된 후보 비표표준지들 중에서 규칙을 만족하는 비교표준지를 선택한다.

⑤ ①~④의 과정을 모든 개별필지들을 대상으로 반복적으로 수행한다.

2. 결과 분석 및 고찰

1) 결과 분석

의사결정트리에 의해 학습된 비교표준지 선정 규칙을 적용하여 비교 표준지를 선정한 결과 표 1과 같이 삼성동의 제1종 전용주거지역에서는 현행 비교표준지와 총 218필지 중에서 142필지가 일치하여 65.13%의 부합률을 나타

```
If Distance <= 86 Then
        If Diff_Jiga <= 120000 Then
                If Distance <= 47 Then
                        "YES"
                Else
                        If Diff_Jiga = 0 Then
                                If Arod <= 0.96 Then
                                        "NO"
                                Else
                                        "YES"
                        Else
                                ...
        Else
                "NO"
Else
        "NO"
```

FIGURE 3. Extracted rule from decision tree

내었으며 일반상업지역의 경우는 총 368필지 중 275필지가 일치하여 74.72%의 부합률을 보이는 것으로 나타났다.

대치동의 경우 제1종 전용주거지역에서는 현행 비교표준지와 총 76필지 중에서 60필지가 일치하여 78.94%의 부합률을 나타내었고, 일반 상업지역은 총 930필지 중 686필지가 일치하여 77.98%의 부합률을 보이는 것으로 나타났다. 연구대상지역 전체 필지에 대하여 현행 비교표준지와의 부합률을 보면, 총 930필지 중에서 686필지가 일치하여 73.76%의 부합률을 나타내었다.

결과를 보면 학습된 비교표준지 선정규칙에 의하여 개별필지의 비교표준지를 선정할 때, 비교표준지가 선정이 되지 않는 필지가 평균적으로 18% 정도 존재한다. 이는 학습된 규칙이 대상지역 전체의 패턴을 나타내는 일반적인 규칙이기 때문이며, 현재 사용되는 비교표준지가

TABLE 1. Coincidence ratio between current comparative standard land parcels and selected comparative standard land parcels using learned rule

동	용도지역	대상필지	선정필지	비선정률	현행 비교표준지와 부합하는 필지	부합률
삼성동	제1종전용 주거지역	218	178	18.34%	142	65.13% (142/218)
	일반상업지역	368	300	18.47%	275	74.72% (275/368)
대치동	제1종전용 주거지역	76	71	6.5%	60	78.94% (60/76)
	일반상업지역	268	214	20.14%	209	77.98% (209/268)
합 계				17.95%		73.76% (686/930)

항상 일정 규칙에 의해서 선정되지 않고 있음을 말해준다. 앞의 현행 비교표준지 선정방법과 그 문제점에서 살펴보았듯이, 비교표준지를 선정하는 기준 자체가 모호성을 띄고 있고 수작업 방식에 의에 발생할 수 있는 오류에 기인한다고 분석된다.

표 2는 동별/용도지역별로 의사결정트리를 구축하는데 있어서 루트노드로 선택되는 속성들을 결정하기 위해 각 속성의 정보 이득량을 계산한 결과이다. 이미 살펴보았듯이 정보이득의 개념은 엔트로피(복잡도)의 감소량으로서, 데이터 셋에서 하나의 속성 A를 선택하여 그 속성 값에 따라 분류하였을 때 분류하고자 하는 클래스(여기서는 실제표준지 여부)가 하나의 클래스로 통일되게 나타나는 경우, 속성 A를 선택함으로서 엔트로피, 즉 복잡도가 감소하는 것이며, 속성 A는 우선적으로 분류의 기준이 되는 것이다. 따라서 이 정보 이득량을 파악함으로써 비교표준지를 선정하는데 있어서 우선적으로 고려하는 사항이 무엇이었는지 분석해 낼 수 있다.

표 2를 보면, 비교표준지를 선정하는 기준을 유추해 볼 수 있다. 실험을 수행한 4개의 사례에 있어서 정보이득이 가장 큰 두 개의 속성을 확인할 수 있다. 이 두 속성은 개별필지와 후보비교표준지들 간의 '지가차이'와 '거리차이'로서 이러한 결과는 현행 비교표준지 선정요령에서 제시한 '유사가격권'과 '인근의 표준지'를 충실히 반영하고 있음을 뜻한다. 또한 토지이용의 차이를 나타내는 LAND_USE 속성은 정보 이득량이 0인 것으로 나타나는데, 이는 연구대상지역의 토지이용(주용도)이 모두 동일하기 때문에 개별필지와 비교표준지사이에 그 차이가 없기 때문이다. 즉 차이가 모두 동일하기 때문에 LAND_USE 속성은 분류를 하는데 있어 전혀 불필요한 속성인 것이다. LAND_USE 속성이 0.237인 대치동 '제1종 전용주거지역'의 경우는 토지이용상황이 주거용과 상업·업무용 두 가지 경우이므로 분류를 위한 속성으로 사용되었다.

각 동의 일반상업지역에 있어서, 전용주거지역과 달리 도로접면 조건의 차이를 나타내는 AROD 속성이 정보이득을 갖는다는 사실을 알 수 있다. 이는 상업지역에서 비교표준지를 선정하는데 있어서 도로접면 조건이 주요한 고려대상의 하나라는 사실을 알 수 있다.

2) 고찰

여기에서는 실험결과와 분석을 토대로 본 연구의 결과를 고찰해 보고 비교표준지 선정업무에 활용할 수 있는 방안을 제시하고자 한다. 실험지역의 경우 학습된 규칙을 이용하여 비교표준지를 선정하였을 때 평균 70%의 부합율을 보이고 있으며, 결과분석을 통해 알 수 있듯이 학습된 규칙은 비교표준지 선정대상 필지들의 용도지역이나 토지이용상황 등에 따라 달라진다. 정보 이득량을 통한 분석을 나타낸 표 2에서는 지역적인 특성, 즉 조사대상지역의 공간적 범위/토지이용상황에 따라 비교표준지를 선정하는데 고려되는 토지특성이 조금씩 다르지만 전체적으로는 유사가격권내의 표준지 또는 인근의 표준지를 선택한다는 패턴을 발견할 수

TABLE 2. Information gain from root node

동	용도지역	정 보 이 득			
		DIFF_JIGA	DISTANCE	AROD	LAND_USE
삼성동	제1종전용 주거지역	0.0469	0.258	0	0
	일반상업지역	0.1321	0.1351	0.0789	0
대치동	제1종전용 주거지역	0.459	0	0	0.237
	일반상업지역	0.1679	0.1202	0.099	0

있었다.

본 연구에서는 기존의 데이터로부터 비교표준지를 선정하는 규칙을 학습하여 애매모호한 기존의 비교표준지 선정 요령을 정형화된 규칙의 형태로 전환함으로써, 비교표준지 선정 시 작업의 객관성과 자동화된 방법을 사용함으로써 효율성을 확보할 수 있었다. 이는 결국 기존의 수작업 방식에 의해 비교표준지를 선정하는 경우 발생하는 주관성과 비효율성을 해결할 수 있는 하나의 방안이 된다. 결과분석의 표 1을 보면, 학습된 비교표준지 선정규칙을 적용하였을 때 전체 대상필지의 17% 가량이 비교표준지가 선정이 되어있지 않으며, 70% 가량의 필지는 기존 수작업 방식에 의한 결과와 일치하고 있다. 그러나 비교표준지가 선정이 된 필지만을 대상으로 한 현행 비교표준지와의 부합율은 90% 정도로서, 비교적 정확히 학습이 되었다고 판단된다. 따라서 연구의 결과를 실제 비교표준지 선정업무에 활용하기 위해서는 1차로 자동선정을 수행하고 비교표준지가 선정되지 않은 필지들은 수작업을 통해 선정하거나 향후 연구를 통해 이러한 문제점을 보완할 수 있는 방법을 도출해야 할 것이다.

결 론

본 연구에서는 기계학습 알고리즘의 하나인 의사결정트리(C4.5 알고리즘)를 사용하여 비교표준지를 선정하는 일반적인 규칙을 학습한 후, 이를 이용하여 비교표준지를 선정하였고 다음과 같은 결과를 얻었다.

연구대상지역인 삼성동의 경우, 제1종 전용주거지역에서 65.13%, 일반상업지역에서 74.72%가 현행비교표준지와 일치하였고, 대치동의 경우 제1종 전용주거지역에서 78.94%, 일반상업지역에서 77.98%가 현행 비교표준지와 일치하였다. 전체적으로는 현행 비교표준지와 73.76%가 일치하여 비교표준지 선정에 대한 자동화의 가능성을 확인할 수 있었다.

자동 선정한 비교표준지가 현행 비교표준지와의 부합률이 대략 70% 정도이다. 학습된 규칙에 의해 선정된 비교표준지와 전통적인 수작업 방식에 의에 선정된 비교표준지의 일치정도를 비교한 것이므로, 실제 인간의 시각(human vision)에 의하여 선정되고 있는 작업을 컴퓨터에 의해 학습된 규칙으로 자동선정하였다는 것을 고려하면 결코 낮은 정확도는 아닐 것이다. 또한 기존의 데이터로부터 비교표준지를 선정하는 규칙을 학습하여 애매모호한 기존의 비교표준지 선정요령을 정형화된 규칙의 형태로 전환하여 자동선정함으로써 비교표준지 선정작업의 객관성과 효율성을 확보할 수 있었다.

연구의 결과를 매년 실시되는 개별공시지가 업무에 활용하기 위해서는 단계적인 방안이 필요하다. 동별-용도지역별로 비교표준지 선정규칙(예; 대치동-일반상업지역)을 추출한 후, 이를 모듈화시켜 비교표준지 자동선정 시스템에서 재사용할 수 있도록 해야 할 것이다. 우선 규칙을 이용하여 자동으로 비교표준지를 선정하고 비교표준지가 선정되지 않는 필지에 대해서는 기존의 수작업 방식으로 선정한다. 이를 반복적으로 적용하게 되면 비교표준지는 더욱더 일정한 규칙을 따라 선정이 될 것이다. 이는 신뢰할 수 있는 비교표준지 선정 규칙이 도출됨을 의미하며, 점진적으로 수작업 비율을 감소시킬 수 있으므로 완전 자동화 시스템에 더욱 근접할 수 있을 것으로 예상되며, 이러한 문제를 보완할 수 있는 방법에 대한 연구가 반드시 필요하다. 또한 본 연구에서는 비교표준지 선정요령에 따라 제한적인 토지특성 만을 고려하였지만, 필지의 향, 형상, 고도와 같은 더욱 다양한 토지특성과 사회적인 요소들도 정형화하여 학습과정에 속성데이터로 사용하는 시도가 향후 연구에 추가되어야 할 것이다.

결론적으로 본 연구에서 제시한 비교표준지 선정방법은 현행 공시지가 산정 절차와 부합하면서 현행 비교표준지 선정 작업의 효율성과

객관성을 개선시킬 수 있었으며, 비교표준지 자동 선정의 관점에서 볼 때 완전 자동화를 위한 초석이 될 수 있을 것이다. KAGIS

참고문헌

건설교통부. 2001. 2002년도 적용 개별공시지가 조사산정지침. 101-104쪽.

김준회. 2001. 데이터마이닝 기법을 이용한 이동통신시장에서의 고객분석에 관한 연구. 인하대학교 석사학위논문. 18-19쪽.

류광렬. 2002. 기계학습. 부산대학교. 17-26쪽 (미발간자료).

문태헌. 2000. GIS 기반 지가산정 및 시뮬레이션 시스템. 한국지리정보학회지 3(2):1-10.

안지현. 2002. 데이터 마이닝을 이용한 병원이용 고객 세분화 분석. 서울대학교 석사학위논문. 5쪽.

장남식, 홍성완, 장재호. 2002. 데이터마이닝. 대청미디어, 서울. 54-59쪽.

정수연. 2002. 표준지 선정 및 분포기준의 문제점과 개선방안. 감정평가 48호

정 현. 1999. Data mining 기법을 이용한 인공신경망 입력변수 선정에 관한 연구 - 기업도산예측모형에의 적용. 서울대학교 석사학위논문. 17쪽.

채미옥. 1997. 서울시 지가의 공간적 분포특성과 지가형성요인에 관한 연구. 서울시립대학교 박사학위논문. 4-9쪽.

채미옥, 권태형. 1997. 공시지가의 균형성 제고방안. 국토개발연구원. 35-38쪽.

최양춘. 2001. 공시지가제도에 관한 연구. 동의대학교 석사학위논문. 63-64쪽.

하선영. 2001. 데이터마이닝을 위한 베이지안 망 구조 학습. 서울대학교 석사학위논문. 7-9쪽.

하철영, 박정호. 2000. 공시지가제도의 개선방안. 동의대학교 동의법정 16:105-119.

홍길순. 1998. 개별공시지가 제도의 발전방향에 관한 연구. 중앙대학교 석사학위논문. 5-6쪽.

Berry, M.J.A. and G. Linoff. 1997. Data Mining Techniques: For Marketing, Sales, and Customer Support. John Wiley & Sons. pp.244-245.

Heckerman, D. 1996. Bayesian Networks for Knowledge Discovery. In: Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press. pp.273-305.

Han, J. and M. Kamber. 2001. Data Mining Concepts and Techniques. Morgan Kaufmann Publishers. pp.196-204.

Holte, R.C. 1993. Very simple classification rules perfrom well on most commonly used datasets. Machine Learning 11(1):63-90. KAGIS.