pISSN : 2508-3384 eISSN : 2508-3392
https://doi.org/10.22640/lxsiri.2016.46.2.43
Journal of Cadastre & Land InformatiX Vol.46 No.2 (2016)
한국국토정보공사「지적과 국토정보」제46권 제2호 2016년 12월 p. 43-55

거리척도와 앙상블 기법을 활용한 지가 추정

Estimating Farmland Prices Using Distance Metrics and an Ensemble Technique

이창로* · 박기호**
Lee, Chang-Ro · Park, Key-Ho

Abstract

This study estimated land prices using instance-based learning. A k-nearest neighbor method was utilized among various instance-based learning methods, and the 10 distance metrics including Euclidean distance were calculated in k-nearest neighbor estimation. One distance metric prediction which shows the best predictive performance would be normally chosen as final estimate out of 10 distance metric predictions. In contrast to this practice, an ensemble technique which combines multiple predictions to obtain better performance was applied in this study. We applied the gradient boosting algorithm, a sort of residual-fitting model to our data in ensemble combining. Sales price data of farm lands in Haenam-gun, Jeolla Province were used to demonstrate advantages of instance-based learning as well as an ensemble technique. The result showed that the ensemble prediction was more accurate than previous 10 distance metric predictions.

Keywords: Instance-based learning, K-nearest neighbor method, Distance metric, Ensemble, Gradient boosting

1. 서 론

자산의 경제적 가치를 판정하고 이를 화폐액으로 표시하는 것을 가치평가(valuation)라 하며, 자산 중에서도 금융자산을 제외한 토지 등의 가치를 판정하는 경우 감정평가(鑑定評價)라는 용어를 일반적으로 사용한다(장희순·방경식 2014). 토지에 대한 감정평가는 통상 다음과 같은 절차를 통하여 이루어진다. 먼저 평가 대상 토지가 소재한 지역의 전체 거래사례를 수집한다. 다음으로 수집한 거래사례 중 평가 대상 토지와 유사한 속성(용도지역, 접면도로, 면적 등)을 가진 거래사례를 선별한다(통상 1개에서 5개 정도를 선별). 마지막으로 선별된 거래사례의 매매가격을 기초로 평가 대상 토지의 가격을 결정한다[1]. 이와 같은 지가추정 과정은 사례 기반 학습(instance-based learning)의 논리와 동일하다. 사례 기반 학습의 경우 먼저 훈련

* 서울대학교 국토문제연구소 연구원 Researcher, Institute for Korean Regional Studies, Seoul National University (First author: spatialstat@naver.com)
** 서울대학교 지리학과 교수 · 국토문제연구소 겸무 연구원 Professor, Department of Geography, Seoul National University, Researcher, Institute for Korean Regional Studies (Corresponding author: khp@snu.ac.kr)

데이터(training data)를 가능한 한 풍부하게 수집한다. 다음으로 수집한 데이터 중 예측하고자 하는 대상과 유사한 속성을 가진 데이터를 선별한 후, 이렇게 선별된 유사 데이터의 값을 기초로 대상의 목표값(target value)을 예측한다.

거래가 일정 수준 이상 이루어지고 이러한 거래사례를 비교적 쉽게 확보할 수 있는 경우 사례 기반 학습은 객관적이고 설득력 있는 토지 감정평가 기법이 된다. 이와 같은 이유로 사례 기반 학습은 토지에 대한 가격 추정시 주된 평가방법으로 널리 사용되어 왔다. 감정평가 분야에서는 사례 기반 학습을 통해 지가를 추정하는 평가방법을 특히 '거래사례 비교법'이라고 칭한다.

목표값 예측에 있어 사례 기반 학습과 대(對)를 이루는 방법이 모형 기반 학습(model-based learning)이다. 모형 기반 학습은 수집한 훈련 데이터 중에서 일부만 선별하여 사용하는 것이 아니고, 전체를 활용하여 일반화된 구조(generalized structure)를 만들어 낸 후 해당 구조에 신규 관찰치를 대입하여 목표값을 예측한다(Quinlan 1993). 선형회귀분석(linear regression analysis)을 통해 모형을 구축한 후, 해당 모형을 이용하여 지가를 예측하였다면 이는 모형 기반 학습을 통해 예측한 셈이 된다. 그러나 이러한 모형 기반 학습은 지가 추정시 사례 기반 학습만큼 널리 활용되지는 않고 있는데, 가치평가 전문가인 감정평가사들은 정확성 등의 이유로 사례 기반 학습을 폭넓게 활용하는 반면, 모형 기반 학습은 공시지가 산정시에만 제한적으로 사용되고 있는 실정이다.

본 연구에서는 지가 추정시 폭넓게 활용되는 사례 기반 학습의 활용에 초점을 맞춘다. 사례 기반 학습은 훈련 데이터와 예측 대상과의 유사성(similarity)을 어떻게 정의할지가 관건인데, 보통 거리척도(distance metric)를 활용하여 거리값이 작게 나오면 유사한 것으로, 크게 산출되면 유사성이 떨어지는 것으로 판단한다. 대표적인 거리척도가 바로 유클리디안 거리(Euclidena distance)라 할 수 있다. 이와 같은 거리척도는 데이터의 특징, 연구 맥락, 연구자의 경험과 판단 등에 따라 매우 다양하게 만들어 낼 수 있다. 본 연구에서는 동일 필지에 대해 다양한 거리척도를 활용하여 복수의 지가를 예측한 후, 이러한 예측값들을 적정하게 결합하는 앙상블 기법(ensemble technique)을 활용하여 최종 지가를 결정한다. 개인보다는 다수의 지성이 더 큰 힘을 발휘하듯, 단일 예측값이 아닌 여러 예측값들을 산출한 후 이를 결합하여 최종값을 결정하는 앙상블 기법은 발표 즉시 많은 관심을 받으며 다양한 분야에서 분석의 대상이 되어 왔다(Banfield 2007; Wang 2008).

모형 기반 학습이 데이터 전체를 활용하는 전역적 접근(global approach)이라면, 거리척도를 활용한 사례 기반 학습은 국지적 접근(local approach)이라 할 수 있다. 국지적 접근은 일반 재화와 달리 표준화시키기 어렵고 개별성이 강한 토지와 같은 자산의 가격을 추정하는데 적합하다. 토지에 대한 감정평가시 거래사례비교법 같은 국지적 방법이 주로 활용되는 이유도 여기에 있다 할 것이다.

또한 감정평가 실무시 한 개의 거래사례에 기반한 가격 산출은 드문 편이고, 통상 여러 개 거래사례에 기초하여 일차적인 가격을 산출한 후('시산가격'이라 한다), 이러한 복수의 가격들을 종합 가감하여 최종 가격을 산출하게 된다. 그러나 시산가격 조정에 있어 감정평가 실무상 명확한 기준이 없어 평균가격으로 최종가격을 정하는 것이 일반적이다. 따라서 복수의 지가를 최적의 상태로 결합하는 앙상블 기법은 시산가격 조정 과정을 보다 효율화함으로써 지가의 정확성을 높일 수 있을 것으로 기대된다.

본 연구에서는 거리척도를 활용한 사례 기반 학습과 앙상블 기법에 대해 살펴보고, 이러한 기법들이 지가 추정시 갖는 타당성과 우수성에 대해 실증 사례를 통해 설명하고자 한다.

2. 거리척도를 활용한 사례 기반 학습

2.1. 거리척도(distance metric)

비교하고자 하는 두 개 관찰치의 유사성, 즉 거리를 측정할 때 가장 널리 알려진 거리척도가 아래의 유클리디안 거리이다.

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^{p}(y_{1j} - y_{2j})^2} \qquad (1)$$

식(1)에서 두 관찰치 x_1, x_2 간의 유클리디안 거리는 관찰치가 가지는 p개 속성값(y_j)의 차이를 제곱하여 합산한 후 제곱근을 적용하여 산출한다. 서론에서 언급한 바와 같이 이러한 거리척도는 무수히 많고, 또 연구자마다 연구 맥락에 맞게 적당한 척도를 개발하여 적용하기도 한다. 따라서 본 연구에서는 발표된 모든 거리척도를 검토하기보다는 문헌에서 비교적 자주 언급되는 10개의 대표적인 거리척도를 활용하고자 한다(Legendre and Legendre 2012). 본 연구에서 활용한 거리척도의 공식과 특징은 Table 1과 같다. Table 1에서 ① ~ ④는 처음부터 거리척도로 개발된 것이며, ⑤ ~ ⑩은 유사성 지수(similarity index)로 개발된 척도이므로 1에서부터 차감하여 거리척도로 변환한 것이다.

Table 1에서 제시된 거리척도를 활용하여 예측하고자 하는 대상과의 거리값을 계산하였다면 거리값이 가까운 사례들을 선별한 후, 아래의 k-최근린법(k-nearest neighbor method)을 이용하여 대상의 목표값을 예측한다.

$$\hat{Y} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum_{i=1}^{k} w_i} \qquad (2)$$

식(2)는 전형적인 사례 기반 학습의 예를 보여주며, 통상 1개 내지 5개 정도(k=1~5)의 유사 사례들을 선별한 후 이들의 값을 산술평균하거나($w_i = 1$), 거리값에 따라 가중치를 달리 부여하여 대상의 값을 예측하게 된다. 이러한 사례 기반 학습은 ① 논리를 이해하기 쉬워 어떠한 속성이 유용한지 또는 불필요한지, 문제점이 무엇인지 비교적 쉽게 파악할 수 있고, ② 데이터의 잡음(noise), 부적절한 속성 등에 큰 영향을 받지 않는 강건한 알고리듬으로 발전시킬 수 있다(Aha et al. 1991).

2.2. 적용 사례

k-최근린법을 사용한 사례 기반 학습은 목표값이 이진변수인 분류(classification)의 문제를 처리하는데 오랫동안 사용되었다. 즉 분류 또는 패턴의 파악(pattern recognition) 문제에 있어 k-최근린법은 가장 직관적이고 전통적인 방법이었다. 그러나 연구자의 경험이나 사전 지식과 효과적으로 결합하여 사용할 경우, 최근에 개발된 복잡하고 정교한 모형과 여전히 견줄 수 있을 정도의 예측 성능을 보여주고 있다(Weinberger et al. 2009). 따라서 안면 인식, 텍스트의 분류, 단백질 타입의 분류 등 여러 분야에서 현재에도 널리 활용되고 있다(Chopra et al. 2005; Liao and Vemuri 2002; Shen and Chou 2005).

k-최근린법을 본 연구에서처럼 목표값이 연속형인 경우 적용한 예도 많이 발견되는데, 도로구간의 교통속도 예측, 고객의 선호도 평가, 코스피 선물지수의 예측 등이 대표적인 예라 할 수 있다(이석준·김선옥 2007; 김명현 외 2015; Rasyidi et al. 2014). 본 연구도 이와 같은 선행연구의 예를 따라 지가, 보다 구체적으로는 농지의 가격을 예측하기 위해 k-최근린법을 활용하였다.

Table 1. Formula and characteristics of distance metrics Source: Legendre and Legendre 2012

Type	Formula	Characteristics
① Euclidean	$\sqrt{\sum_{j=1}^{p}(y_{1j}-y_{2j})^2}$	-
② Manhattan	$\sum_{j=1}^{p}\lvert y_{1j}-y_{2j}\rvert$	택시의 주행경로와 유사하다고 하여 taxicab metric이라 고도 불림 It is like the distance travelled by a taxicab around blocks in a city, and hence also called a taxicab metric.
③ Canberra	$\sum_{j=1}^{p}\left[\frac{\lvert y_{1j}-y_{2j}\rvert}{(y_{1j}+y_{2j})}\right]$	두 관찰치가 모두 0의 값을 갖는 경우는 공식 수정 필요 It needs to be modified when both observations have value of zero.
④ Chord	$\sqrt{\sum_{j=1}^{p}\left[\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p}y_{1i}^2}}-\frac{y_{2j}}{\sqrt{\sum_{j=1}^{p}y_{2i}^2}}\right]^2}$	유클리디안 거리를 표준화한 개념 It is a sort of normalized Euclidean distance.
⑤ Hamman	$1-\frac{a+d-b-c}{p}$	a: 두 관찰치가 모두 보유한 속성의 수. The number of features coding the two observations 1. d: 두 관찰치가 모두 미보유한 속성의 수. The number of features coding the two observations 0. b: x_1은 보유, x_2는 미보유한 속성의 수. The number of features for which one observation coding 1 and the other one coding 0. c: x_1은 미보유, x_2는 보유한 속성의 수. The number of features for which one observation coding 0 and the other one coding 1. p = a + b + c + d
⑥ Jaccard	$1-\frac{a}{a+b+c}$	
⑦ Podani	$1-2\times\frac{a-b+c-d}{p(p-1)}$	
⑧ Soergel	$1-\frac{b+c}{b+c+d}$	
⑨ Russel	$1-\frac{a}{p}$	
⑩ Gower	$1-\frac{1}{p}\sum_{j=1}^{p}s_{12j}$	$s_{12j}=1$: 두 관찰치가 동일한 속성을 보유. The number of features for which the two observations are coded identically. $s_{12j}=0$: 두 관찰치가 상이한 속성을 보유. The number of features for which the two observations are coded differently.

3. 여러 예측치의 병합

3.1. 앙상블 기법

여러 사람들의 추측이나 의견을 합치면 소수 전문가 그룹의 예측 결과보다 나을 수 있다. 즉 목표값을 예측하고자 할 때 가장 우수한 것으로 확인된 모형의 단일 추정값을 사용하는 것이 아니라, 여러 개 모형에서 산출된 추정값들을 병합하여 최적의 추정값을 산출하는 과정을 기계학습 분야에서 앙상블 기법이라고 한다. 복수의 학습결과를 병합하여 결과적으로 보다 나은 성능을 내고자 하는 앙상블 기법은 예측 분야에서 특히 각광받고 있다(Schapire 1999).

이와 같은 앙상블 개념을 실행할 수 있는 앨고리듬은 다양한데, 일반적으로 접하게 되는 앨고리듬이 배깅(bagging)과 부스팅(boosting)이다. 배깅은 데이터로부터 일부 데이터를 복원 추출하여, 즉 부트스트랩(bootstrap)을 통해 여러 개(예를 들어 100개)의 부분 데이터 집합(subset data)을 만들어 내고, 이러한 부분 데이터 집합에 모형을 100번 적합하여 예측값을 각각 계산한다. 마지막으로 이러한 예측값을 평균하여 최종 예측치를 정하게 된다.

반면 부스팅은 부트스트랩에 기초한 여러 개의 부분 데이터 집합을 생성하지 않는다. 대신 최초의 원데이터를 계속하여 수정하면서 모형을 업데이트한다. 즉 최초의 모형을 구성한 후, 종속변수 Y가 아닌 잔차를 업데이트하는 방식으로 모형을 수정하게 된다. 배깅의 경우 선형회귀모형 등 전통적 모형보다 예측 성능이 뛰어남을 보여 준 사례도 존재하지만(Fanelli et al. 2013), 여러 예측값을 병합할 때 단순히 평균값을 적용하는데 그치는 한계가 있다. 반면 부스팅은 좀더 효과적인 예측값 병합과정을 거치는데, 직전 과정에서 산출된 잔차, 즉 오차를 계속해서 줄여나가는 방향으로 예측값 수정을 한다. 본 연구에서는 이러한 특징을 고려하여 부스팅 앨고리듬을 적용하였다.

부스팅 앨고리듬 역시 세부 실행방법에는 여러 가지가 있는데(Gradient Boosting, AdaBoost, XGBoost, Gentle Boost 등), 이들 세부 실행방법 간에 근본적인 차이는 없으며 적용 과정상에 약간의 상이점만 있을 뿐이다. 본 연구에서 적용한 부스팅 앨고리듬은 Friedman(2001)의 경사 부스팅(Gradient Boosting)을 따랐다. 경사 부스팅의 실행 논리에 대한 개념적 설명은 다음과 같다[2]..

먼저 종속변수 Y에 대해 다음과 같이 초기 모형을 설정한다.

$$Y = f(x) + \epsilon \tag{3}$$

상기와 같은 초기 모형에서 만약 오차항 ϵ이 순수 오차항, 즉 백색 잡음(white noise)이 아니고 종속변수 Y와 어떠한 상관성(correlation)을 갖는 변수라면 오차항에 대해 다음과 같은 두 번째 모형을 설정할 수 있다.

$$\epsilon = g(x) + \epsilon_2 \tag{4}$$

이 경우 오차항 ϵ_2의 크기는 ϵ의 크기보다 작아질 것이다. 동일한 논리로 ϵ_2가 순수 오차항이 아니라면 다음과 같은 세 번째 모형을 설정할 수 있다.

$$\epsilon_2 = h(x) + \epsilon_3 \tag{5}$$

마찬가지로 오차항 ϵ_3의 크기는 ϵ_2의 크기보다 작아질 것이다. 지금까지의 과정을 하나의 식으로 표현하면 다음과 같다.

$$Y = f(x) + g(x) + h(x) + \epsilon_3 \tag{6}$$

위 식에서 각 단계에 대한 적절한 가중치를 찾아낼

수 있다면 아래와 같은 모형 구성이 가능할 것이다.

$$Y = \alpha f(x) + \beta g(x) + \gamma h(x) + \epsilon_4 \quad (7)$$

상기와 같은 일련의 식들이 바로 경사 부스팅의 기본적 개념이라 할 수 있다. 오차항은 실지로 순수 오차항일수도 있으나 앙상블이 약한 학습자(weak learner)들을 묶어 하나의 강한 학습자(strong learner)를 만드는 개념이므로 약한 학습자에서 산출된 오차항이 순수 오차항일 가능성은 매우 낮다. 또한 위와 같은 오차항 축소의 과정을 충분히 길게 진행할수록 오차항 자체의 크기는 작아지겠지만 실무상 과다적합(over-fitting)의 위험을 방지하기 위해 적절한 단계에서 진행을 멈추는 것이 일반적이다.

3.2. 적용 사례

비교적 초기에 등장해 광범위하게 활용된 Adaboost (adaptive boosting)는 모형의 매 적합단계에서 오차가 크게 나타난 관찰치에 더 큰 가중치를 부여하여 모형을 업데이트하였다. Park and Bae(2015)의 버지니아주 소재 주택가격 예측, Alfaro et al.(2008)의 기업 파산 확률 예측 등이 Adaboost를 활용한 사례에 해당된다.

이후 Adaboost 앨고리듬을 지수 손실함수(exponential loss function)에 대한 경사 강하(gradient descent) 방법으로 접근하는 연구가 진행되었고, 이러한 연구 흐름 속에서 Friedman(2001)은 여러 가지 손실함수에 활용 가능한 부스팅 앨고리듬을 제시하였다. 이렇게 제시된 앨고리듬 중의 하나가 경사 부스팅이며 앞서 설명하였듯이 순차적으로 잔차를 축소시켜 나가는 일종의 잔차 적합(residual algorithm) 앨고리듬이라고 할 수 있다(김희종·김형도 2014). Li et al.(2007)의 시뮬레이션 자료를 대상으로 한 예측 사례, Lemmens and Croux(2006)의 고객의 이탈확률 예측 등이 경사 부스팅을 활용한 예에 해당된다.

4. 사례 분석

4.1. 데이터 설명

본 연구에서 사용한 데이터는 2011년부터 2013년까지 3년간 신고된 전라남도 해남군의 토지(농지) 실거래가 신고자료 4,375건이다[3]. Table 2는 해남군 농지 실거래가 자료의 기초 통계량을 보여준다.

Table 2를 보면 해남군의 전형적인 농지는 농림지역에 속하는 면적 약 2,200㎡ 내외의 토지임을 알 수 있고, 거래가격은 약 23,000원/㎡ 수준으로 형성되었음을 알 수 있다. 본 분석에서는 전체 신고자료 4,375건을 임의분할(random split)하여 첫 50%는 모형 구축을 위한 훈련 데이터(train data)로 사용하였고(2,187건), 나머지 50%는 구축된 모형의 예측 성능을 검토하기 위한 검증 데이터(test data)로 유보하였다

Table 2. Descriptive statistics(n=4,375)

Items	Min.	Mean	Max.	Stan. Dev.
Area(㎡)	6	2,182	47,950	2,546
Price(KRW/㎡)	807	23,040	99,680	21,607

Zone	Residential 13 lots	Industrial 9 lots	Green 142 lots	Management 1,591 lots	Agriculture 2,419 lots	Nature Preservation 201 lots
Use	dry field 2,005 lots (including orchard)			paddy field 2,370 lots		
Sales year	1,530 lots in 2011		1,343 lots in 2012		1,502 lots in 2013	

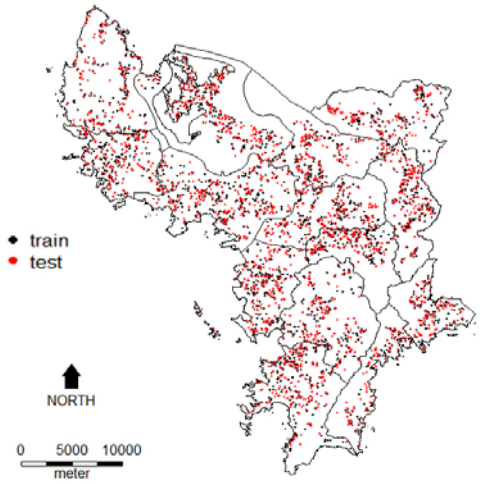


Figure 1. Distribution of training and test data

(2,188건). Figure 1은 훈련 데이터와 검증 데이터로 임의분할된(50:50) 자료의 공간적 분포를 보여주고 있다.

4.2. 농지 간 유사성(거리)의 측정

농지 간 유사성, 즉 거리를 측정하기 위한 항목은 실거래가 신고자료로부터 확인할 수 있는 항목으로 국한하였는데, 다음과 같은 7개 항목을 이용하여 농지 간 거리를 계산하였다.

Table 3에서 읍면동 항목의 경우 혁신도시 등이 들어서는 산이면 및 화원면 일대 농지가격은 타 읍면동 대비 월등히 높은 수준인 바, 별도의 거리 측정 항목으로 포함시켰다. 또한 용도지역의 경우 통상 주거지역 〉 공업지역 〉 녹지지역 〉 관리지역 〉 농림지역 〉 자연환경보전지역 순으로 지가수준이 형성되는 바, 용도지역 항목을 일종의 서열척도(ordinal metric)로 보아 표와 같이 변환하였다. 마지막으로 도로접면 항목은 농지에 접한 도로의 폭이 12m ~ 25m 정도인 경우 중로, 8m ~ 12m인 경우 소로, 8m 미만이면서 자동차 통행이 가능한 경우 세로가, 8m 미만이면서 자동차 통행이 불가능한 경우 세로불, 그리고 접한 도로가 없는 경우 맹지로 분류되어 있었다. 이러한 다항범주 변수를 거리척도 계산에 투입하기 위하여 역시 용도지역 항목과 유사하게 서열척도로 취급하여 표와 같이 변환하였다. 접한 도로의 폭이 커질수록 지가수준이 높아지는 것은 토지 거래시장에서 흔히 관찰되는 일반적인 현상이라 할 수 있다.

Table 3. Items used in calculating distances

Items	Transformation, coding, etc.
Area	Normalized
Sales year	Coded as followings: 2011 = 0, 2012 = 1, 2013 = 2
Slope	level = 0, sloping = 1
Use	dry field = 0, paddy field = 1
District dummy	lots in Sani-myeon and Hwawon-myeon = 1, otherwise = 0
Zone	residential = 6, industrial = 5, green = 4, management = 3, agriculture = 2, nature preservation = 1
Adjacent road	road width between 12 and 25 meters = 5, between 8 and 12 meters = 4, less than 8 meters = 3, less than 8 meters and vehicle inaccessible = 2, no road =1

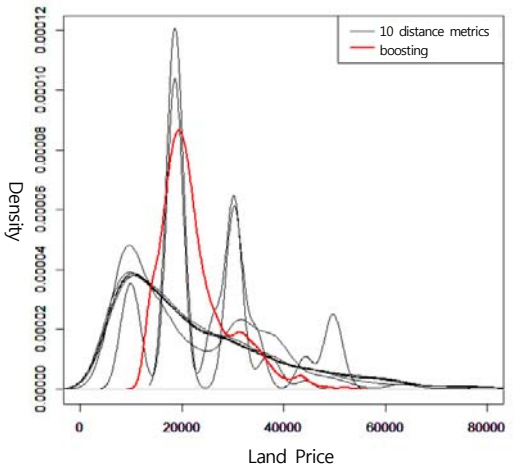


Figure 2. Distribution of estimates from the 10 distance metrics and boosting

4.3. 예측 성능의 비교

본 연구에서는 Table 3의 7개 항목을 사용하여 Table 1에 제시된 10개의 거리 척도를 계산하였다. 이와 같이 계산된 거리값을 토대로 k-최근린법을 적용하여 유보한 검증 데이터 2,188건의 지가를 예측하였다 4). 또한 마지막으로 10개의 거리척도에 의해 계산된 예측값들을 경사 부스팅 앨고리듬에 투입하여 앙상블 예측값을 산출하였다. Figure 2는 검증 데이터를 대상으로 한 추정한 11개 예측값들의 분포 현황을 보여준다. Figure 2를 보면 경사 부스팅에 의한 예측값 분포가 나머지 10개 거리척도에 의한 예측값들 분포의 중간 정도에 위치하며, 경사 부스팅 최빈값(mode)이 Table 2 기초 통계량에서 계산된 평균값 23,000원/㎡과 유사한 20,000원/㎡ 수준에서 형성되어 있음을 알 수 있다.

다양한 거리척도를 활용하여 추정된 10개의 예측값들 상호 간의 정확성 비교와 이러한 값들을 토대로 산출된 부스팅 앙상블값의 예측 성능은 다음과 같은 2가지 지표를 사용하여 비교하였다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y_i} - y_i\right)^2} \tag{8}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y_i} - y_i}{y_i}\right| \tag{9}$$

RMSE(Root Mean Squared Error)와 MAPE(Mean Absolute Percent Error)는 모두 실제값과 예측값의 차이를 가늠하는 지표로서 예측의 정확도를 비교할

Table 4. RMSE

거리척도	RMSE
Gower	24,439
Hamman	22,519
Jaccard	22,518
Canberra	24,685
Russel	22,873
Podani	24,019
Soergel	24,580
Euclidean	24,548
Manhattan	24,564
Chord	24,380
Boosting	21,342

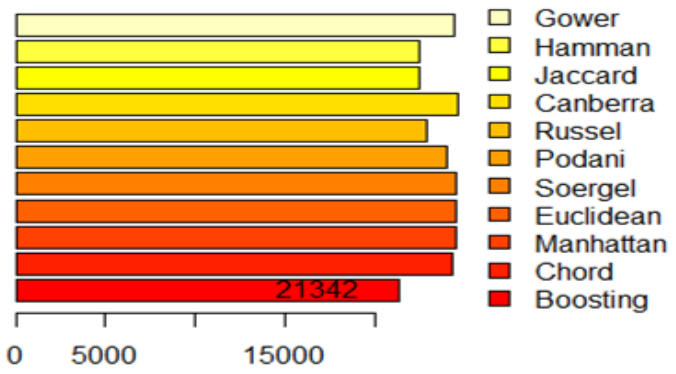
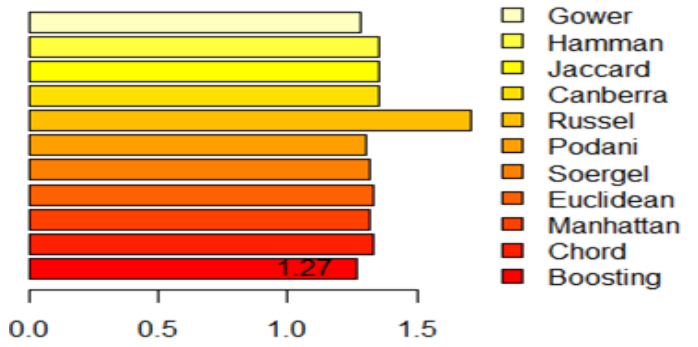


Table 5. MAPE

거리척도	RMSE
Gower	1.28
Hamman	1.35
Jaccard	1.35
Canberra	1.35
Russel	1.71
Podani	1.30
Soergel	1.32
Euclidean	1.33
Manhattan	1.32
Chord	1.33
Boosting	1.27



때 자주 사용된다. 검증 데이터를 대상으로 한 10개 거리척도 및 앙상블에 의한 예측 정확도는 Table 4 및 Table 5와 같다. 거리척도에 의한 10 예측값 사이에 큰 격차는 없는 것으로 보이며, 다만 RMSE 기준으로 Jaccard 예측값이, MAPE 기준으로 Gower 예측값이 근소하나마 가장 우수한 성능을 보이는 것으로 나타났다. 그러나 최종적으로 경사 부스팅을 적용한 앙상블 예측값이 RMSE 및 MAPE 지표 모두에서 기존 10개의 거리척도보다 예측 성능이 우수한 것으로 산출되었다.

부스팅 같은 앙상블 기법을 적용한다 하더라도, 상황에 따라 예측 성능은 기존 예측값들 대비 전혀 개선되지 않을 수 있다. 예측값의 다양성, 즉 예측값 사이의 낮은 상관성(correlation)은 앙상블 예측 성능 향상의 관건이라고 할 수 있다(Gama et al., 2005, p.406). 예를 들어 앙상블 기법에 투입된 기존 예측값들 사이의 상관계수(pearson coefficient)가 0.80을 초과한다면 앙상블 기법을 활용한 결과에 대해 현격한 개선을 기대하기 어려울 수 있다. Table 6은 검증 데이터를 대상으로 추정한 예측값 사이의 상관계수를 보여주는

Table 6. Correlation coefficients between estimates (results from test data)

구 분	G	H	J	C	R	P	S	E	M
Gower									
Hamman	0.12								
Jaccard	0.12	1.00							
Canberra	0.38	0.07	0.07						
Russel	0.28	0.57	0.57	0.23					
Podani	0.21	0.36	0.36	0.23	0.45				
Soergel	0.53	0.07	0.07	0.87	0.20	0.23			
Euclidean	0.52	0.07	0.07	0.86	0.19	0.23	0.99		
Manhattan	0.53	0.07	0.07	0.87	0.20	0.23	1.00	0.99	
Chord	0.51	0.07	0.07	0.85	0.19	0.23	0.96	0.95	0.96

데, 낮게는 0.07에서 높게는 1.00까지 분포하고 있다. 이는 예측값들의 다양성을 나타내는 것으로 해석할 수 있으며, 이러한 다양성을 통해 앙상블 기법에 의한 성능 개선이 이루어졌다고 풀이할 수 있다.

5. 결론

본 연구는 농지 가격을 추정하기 위하여 가치평가 전문가인 감정평가사들의 감정평가 절차와 유사한 사례 기반 학습의 논리를 활용하였다. 즉 사례 기반 학습 중 k-최근린법을 이용하여 전라남도 해남군의 농지 가격을 추정하였다. k-최근린법은 추정하고자 하는 대상과 다른 관찰치들 사이의 유사성, 즉 거리를 어떻게 계산할지가 관건인데 본 연구에서는 문헌에 비교적 자주 등장하는 10개 척도를 사용하여 거리를 계산하였다. 10개 거리척도를 이용하여 추정한 10 종류의 농지 가격 예측값 사이에 현격한 정확도 차이는 발견할 수 없었다. 근소하나마 RMSE 기준으로 Jaccard 거리척도에 의한 예측값이, MAPE 기준으로 Gower 거리척도에 의한 예측값이 비교적 정확하게 지가를 추정한 것으로 분석되었다. 본 연구에서는 이러한 다양한 예측값 사이의 성능 비교에 머물지 않고 이들 예측값들을 병합하는 앙상블 기법의 논리를 적용하여 최종 농지 가격을 결정하였다. 앙상블 기법 중 자주 사용되는 경사 부스팅 앨고리듬을 적용하여 최종 가격을 결정하였고, 그 결과 기존 10개의 예측값들보다 추정의 정확도가 높아진 것을 확인할 수 있었다. 본 연구에서는 기존 10개 예측값들 사이의 유사성이 비교적 낮아서, 즉 일정 수준 다양성을 가지고 있었기에 이러한 예측 정확도의 개선이 가능했던 것으로 풀이하였다.

본 연구는 다음과 같은 세 가지 측면에서 의의를 갖는다. 먼저 선형회귀모형 등 모형 기반 학습이 아닌 감정평가 실무에서 일반적으로 활용되는 거래사례비교법의 절차, 즉 사례 기반 학습의 논리를 활용하여 지가 예측을 시도하였다는 점이다. 감정평가 분야에서 전문가가 가장 빈번하게 활용하는 거래사례비교법은 그 논리가 k-최근린법과 동일하며 따라서 향후 k-최근린법의 적극적 응용이 필요하다. 두 번째로 k-최근린법을 적용하여 대상을 분류하거나 예측하는 경우 거리척도는 통상 유클리디안 거리를 활용하는데 그쳤으나 본 연구에서는 유사성을 측정하는 다양한 거리척도를 소개하는 동시에 실제 지가 예측에 사용하였다는 점에서 타 연구와 차별성을 갖는다. 마지막으로 산출된 여러 개 예측값들 중 비교적 성능이 우수하다고 판단되는 1개 예측값을 선택하는 것이 아니라, 이들 예측값들을 적정하게 병합하는, 앙상블 기법의 논리를 사용하여 최종 예측값을 결정하였으며, 그 결과 예측의 정확도가 추가적으로 개선될 수 있음을 보였다는 점이다.

그러나 본 연구는 보완되거나 발전시킬 여지가 많다. 먼저 앙상블 기법은 예측값들 사이의 유사성이 낮을 때 예측의 정확도가 높아질 수 있다. 즉 앙상블 기법을 사용한다고 하여 항상 정확도가 높아지는 것은 아니므로 데이터가 어떠한 특징을 가질 때 예측 정확도가 크게 개선될 수 있는지, 반대로 그러한 개선효과가 거의 없는지 등을 향후 과제로 살펴볼 필요가 있다.

또한 본 연구에서는 도로접면, 이용상황 등 토지의 물리적 특성을 주된 설명변수로 사용하였다. 즉 물리적 거리를 계산하여 지가를 예측하였으나, 지가 형성에는 이러한 물리적 요인뿐 아니라 주민의 소득 수준, 범죄율, 학군 등 사회경제적 요인도 영향을 주기 마련이다. 특히 농지의 경우라면 농산물 집하 및 운반의 편의성, 인접 대도시까지의 접근성, 주변 지역 성숙에 따른 해당 지역 개발(도시화) 압력 등이 보다 중요한 지가 형성 요인으로 작용할 수 있다. 따라서 이러한 사회경제적 측면의 거리를 측정하여 유사성 계산에 반영한다면 지가 정확도의 추가 개선이 가능할 것으로 보인다.

다음으로 본 연구에서는 7개 토지특성 항목을 기준으로 거리척도를 계산하였는데, 각 항목별 가중치를

별도로 부여하지 않았으므로 7개 항목에 대해 동일한 비중을 주어 거리 계산을 한 셈이다. 그러나 용도지역이나 도로접면은 감정평가 실무에서 특히 중요시하는 항목으로 이러한 항목에 높은 가중치를 두어 거리척도를 계산할 수 있을 것이다.

마지막으로 본 연구는 사례 기반 학습에 초점을 맞추었으나 모형 기반 학습을 병행하여 두 종류의 학습에서 나온 결과물을 모두 고려하여 최종 예측치를 결정할 수도 있을 것이다.

본 연구가 농지 가격 예측을 포함한 사회 여러 분야에 사례 기반 학습 및 앙상블 기법에 대한 관심을 제고하는데 단초가 되기를 기대한다.

주1. 이 때 거래사례와 평가 대상 토지 간의 상이점은 전문가의 판단으로 합리적인 범위 내에서 가감조정한다.

주2. Kuhn and Johnson(2013)의 설명에서 맥락을 약간 변형하여 요약한 것이다.

주3. 국토연구원은 실거래가 기반 부동산 가격공시제도 도입을 위한 기초 연구를 2014년 수행한 바 있으며, 동 연구에서 전형적인 농촌지역으로 전라남도 해남군을 선정하여 실증 분석을 실시하였다. 본 연구에서는 해당 분석에서 사용된 데이터를 사용하였다.

주4. 후술하는 RMSE 및 MAPE가 가장 작아지는 것을 기준으로 최근린 관찰치의 수 k는 3으로 정하였다. 그러나 k=4 또는 k=5의 경우에도 큰 성능의 차이는 없었다. 마지막으로 대상 농지와 가장 유사하다고 선별된 3개의 사례 가격을 기준으로, 거리값의 역수를 가중치로 하여 대상 농지의 가격을 정하였다.

참고문헌
References

김명현, 이세호, 신동훈. 2015. K-Nearest Neighbors (K-NN) 알고리즘을 통한 KOSPI200 선물지수 예측효과 연구. 대한경영학회지. 28(10):2613-2633.

Kim MH, Lee SH, Shin DH. 2015. Predictability Test of K-Nearest Neighbors Algorithm: Application to the KOSPI 200 Futures. *Korea Business Management Journa*l. 28(10):2613-2633.

김희종, 김형도. 2014. 그라디언트 부스팅과 균형 분류를 이용한 채무 불이행 예측. 한국정보기술학회 논문지. 12(1):155-164.

Kim HJ, Kim HD. 2014. Predicting Loan Defaults with Gradient Boosting and Balanced Classification. J*ournal of Advanced Information Technology and Convergence*. 12(1):155-164.

이석준, 김선옥. 2007. 협업필터링에서 고객의 평가치를 이용한 선호도 예측의 사전평가에 관한 연구. Asia Pacific Journal of Information Systems. 17(4):187-206.

Lee SJ, Kim SO. 2007. Pre-evaluation for Prediction Accuracy by Using the Customer's Ratings in Collaborative Filtering. *Asia Pacific Journal of Information Systems*. 17(4):187-206.

장희순, 방경식. 2014. 부동산 용어사전. 부연사.

Jang HS, Bang KS. 2014. *Real Estate Dictionary*. Buyeonsa.

Aha DW, Kibler D, Albert MK. 1991. Instance-based Learning Algorithms. *Machine Learning*. 6(1):37-66.

Alfaro E, García N, Gámez M, Elizondo D. 2008. Bankruptcy Forecasting: An Empirical Comparison of AdaBoost and Neural Networks. *Decision Support Systems*. 45(1):110-122.

Banfield RE. 2007. A Comparison of Decision Tree Ensemble Creation Techniques, IEEE Transactions on Pattern Analysis and Machine Intelligence. 29(1):173-180.

Chopra S, Hadsell R, LeCun Y. 2005. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In: Computer Vision and Pattern Recognition. Proceedings of a Conference Held by IEEE Computer Society; 2005 Jun 20;

San Diego (CA); 2005. Vol. 1. p. 539-546.

Fanelli G, Dantone M, Gall J, Fossati A, Gool L. 2013. Random Forests for Real Time 3D Face Analysis. *International Journal of Computer Vision*. 101(3):437-458.

Friedman JH. 2001. Greedy Function Approximation: a Gradient Boosting Machine. *Annals of Statistics*. 29(5): 1189-1232.

Gama J, Camacho R, Brazdil P, Jorge A, Torgo L. 2005. Machine Learning: ECML 2005. Proceedings of a symposium held at the 16th European Conference on Machine Learning; 2005 Oct 3-7; Porto, Portugal; 2005. p. 601- 608.

Kuhn M, Johnson K. 2013. *Applied Predictive Modeling*. New York: Springer, p. 389-400.

Legendre P, LF Legendre. 2012. *Numerical Ecology*. London: Elsevier, p. 296-298.

Lemmens A, Croux C. 2006. Bagging and Boosting Classification Trees to Predict Churn. *Journal of Marketing Research*. 43(2):276-286.

Li P, Wu Q, Burges CJ. 2007. Mcrank: Learning to Rank Using Multiple Classification and Gradient Boosting. In: Proceedings of a symposium held at the 21st Annual Conference on Neural Information Processing Systems; 2007 Dec 3-5; Vancouver (BC); 2007. p. 897-904.

Liao Y, Vemuri VR. 2002. Use of K-Nearest Neighbor Classifier for Intrusion Detection. *Computers & Security*. 21(5):439-448.

Park B, Bae JK. 2015. Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data. *Expert Systems with Applications*. 42(6):2928-2934.

Quinlan JR. 1993. Combining Instance-based and Model-based Learning. In: Proceedings of a symposium held at the 10th International Conference on Machine Learning; 1993 Jun 27-29; Amherst (MA); 1993. p. 236-243.

Rasyidi MA, Kim J, Ryu KR. 2014. Short-Term Prediction of Vehicle Speed on Main City Roads Using the K-Nearest Neighbor Algorithm. *Journal of Intelligence and Information Systems*. 20(1):121-131.

Schapire RE. 1999. Theoretical Views of Boosting. In: Proceedings of a symposium held at the 4th European Conference, EuroCOLT on Computational Learning Theory; 1999 Mar 29-31; Nordkirchen, Germany; 1999. p. 1-10.

Shen H, Chou KC. 2005. Using Optimized Evidence-Theoretic K-Nearest Neighbor Classifier and Pseudo-amino Acid Composition to Predict Membrane Protein Types. *Biochemical and Biophysical Research Communications*. 334(1):288-292.

Wang YQ. 2008. Building Credit Scoring Systems Based on Support-based Support Vector Machine Ensemble. In: Proceedings of a symposium held at the 4th International Conference on Natural Computation; 2008 Oct 18-20; Jinan, China; 2008. p. 323-326.

Weinberger KQ, Blitzer J, Saul LK. 2009. Distance Metric Learning for Large Margin Nearest Neighbor Classification. J*ournal of Machine Learning Research*. 10: 207-244.

2016년 08월 23일 원고접수(Received)
2016년 10월 27일 1차심사(1st Reviewed)
2016년 12월 07일 게재확정(Accepted)

초 록

본 연구는 사례 기반 학습(instance-based learning)의 논리를 활용하여 지가를 추정하였다. 다양한 사례 기반 학습 기법 중 k-최근린법을 이용하였으며, k-최근린법 적용시 유사성을 측정하는 거리척도는 유클리디안 거리를 비롯해 문헌에 비교적 자주 등장하는 10개의 거리척도를 사용하였다. 본 연구에서는 k-최근린법에 의한 10 종류의 예측값 중 가장 우수한 성능을 보이는 1개의 예측값을 최종 가격으로 선택하는 대신, 이들 예측값들을 병합하는 앙상블(ensemble) 기법의 논리를 적용하여 최종 예측값을 결정하였다. 앙상블 기법 중 일종의 잔차 적합 모형인 경사 부스팅 앨고리듬을 적용하여 최종 가격을 정하였다. 본 연구에서는 이러한 사례 기반 학습과 앙상블 기법의 이점을 실증적으로 제시하기 위해 전라남도 해남군 소재 농지를 사례로 하여 가격을 추정하였으며, k-최근린법에 의한 10 종류의 예측값보다 앙상블 기법에 의한 가격이 보다 정확한 것을 확인할 수 있었다.

주요어 : 사례 기반 학습, k-최근린법, 거리척도, 앙상블, 경사 부스팅