디지털산업정보학회 논문지 제19권 제1호-2023년 3월
http://dx.doi.org/10.17662/ksdim.2023.19.1.109

의사결정나무를 활용한 신경망 모형의 입력특성 선택: 주택가격 추정 사례

윤 한 성*

Decision Tree-Based Feature-Selective Neural Network Model: Case of House Price Estimation

Yoon Han-Seong

〈Abstract〉

Data-based analysis methods have become used more for estimating or predicting housing prices, and neural network models and decision trees in the field of big data are also widely used more and more. Neural network models are often evaluated to be superior to existing statistical models in terms of estimation or prediction accuracy. However, there is ambiguity in determining the input feature of the input layer of the neural network model, that is, the type and number of input features, and decision trees are sometimes used to overcome these disadvantages.

In this paper, we evaluate the existing methods of using decision trees and propose the method of using decision trees to prioritize input feature selection in neural network models. This can be a complementary or combined analysis method of the neural network model and decision tree, and the validity was confirmed by applying the proposed method to house price estimation. Through several comparisons, it has been summarized that the selection of appropriate input characteristics according to priority can increase the estimation power of the model.

Key Words : Neural Network Model, Decision Tree, Input Feature, House Price

Ⅰ. 서론

주택산업의 정보화 및 금융시장 연결성에 따라 부동산 투자자금 및 투자신탁 등의 관련 금융투자 시장이 활성화되고 있다[1]. 이에 따라 정책결정자는 물론 일반 대중들에게도 주택가격 또는 주택가격의 추정은 주요 관심사가 되고 있다[2].

* 경상국립대학교 경영대학 교수

관련 연구들에서 주택가격의 추정(house price estimation)은 주택과 관련되는 특징 또는 경제적 요인들을 주요 데이터로 활용한다[1-5]. 여기서 주택과 관련되는 특징으로는 주택의 연령, 유형, 층수, 부지크기, 건물수, 외관, 위치 등이 포함되며, 경제적 요인으로는 GDP, GNP, 물가지수, 주가지수, 이자율, 부도율, 고용률 등이다. 다양한 데이터들이 분석의 범위 또는 목적에 따라 주택가격 추정에 선택적으로 활용된다.

주택가격 또는 주택가격지수의 추정이나 예측의 분석에 대해 전통적인 계량경제(econometrics) 분야에서는 회귀분석 또는 시계열분석을 중심으로 다양한 파생형(variation)의 방식이 활용되고 있으며[1], 데이터 기반의 여러 분석방식들도 다양하게 적용되고 있다[2-5]. 관련 분야에서 활용되는 빅데이터 분야의 기법으로는 신경망(neural network) 모형, 의사결정나무(decision tree), 랜덤포레스트(random forest), 앙상블(ensemble) 학습 등의 다양한 모형이나 알고리즘을 들 수 있다[2-5]. 그중에서 신경망 모형은 뉴런들이 상호연결되는 인간의 두뇌구조를 모방한 지도학습 방식으로 복잡하고 비선형적인 관계성을 가지는 다변량을 분석할 수 있다. 인공신경망 모형은 추정이나 예측의 정확성에 있어서 기존의 통계적 모형에 비해 우수하게 평가되기도 한다[6].

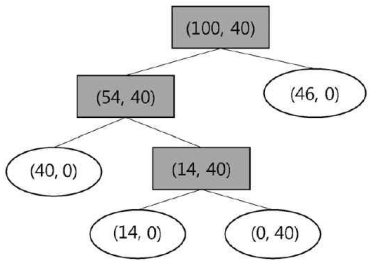
입력층, 은닉층, 출력층의 여러 노드(node)들이 층간에 서로 연결되는 일반적인 신경망 모형은, 훈련데이터(training data)에 의해 학습된 노드(node) 간의 연결 가중치 및 노드별 활성화 함수(activation function)에 의해 계산되는 출력층의 출력값에 의해 추정이나 분류 등을 처리하게 된다. 그런데 신경망 모형의 입력층에 대응하는 입력 데이터, 즉 입력특성의 종류 및 수를 결정하는데 모호성이 존재하며, 이에 대한 체계적인 방법의 결여가 단점으로 지적되기도 한다[7]. 이러한 단점을 극복하는 방안으로 문제영역의 데이터에 대한 의사결정나무를 활용하기도 하지만[6-9], 신경망 모형의 성능을 최적화하는 입력특성을 고려하는 면은 부족하다.

본 논문에서는 신경망 모형의 입력특성(feature) 선택에 의사결정나무를 활용하는 기존의 방안들을 평가하고, 구성하고자 하는 신경망 모형의 성능을 고려한 개선된 방식을 제안하기로 한다. 그리고 제안된 방식을 주택가격 추정문제에 적용하고 성능을 평가하기로 한다. 본 논문에서 정리한 내용은 의사결정나무 및 신경망 모형의 결합 또는 상호보완의 측면에서 중요성과 의미가 있다.

II. 이론적 배경

2.1 의사결정나무와 신경망 모형

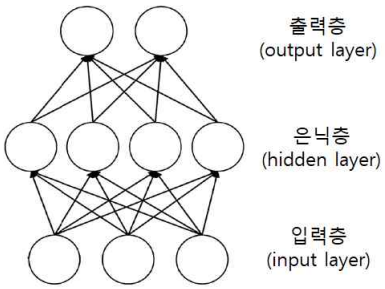
의사결정나무는 표본집단에 대해 선택한 설명변수의 값으로 분리된 그룹별 개체들이 가지는 목표변수 값의 불순도(impurity)를 최소화하는 분지(splitting) 과정을 반복하여 트리(tree) 형태의 분류모형을 구성하는 방식이다. 의사결정나무 분석의 대표적인 알고리즘은 CART(Classification and Regression Trees), CHAID(CHi-square Automatic Interaction Detection), C4.5 등이다. <그림 1>은 두 가지의 목표값을 각각 가지는 100개 및 40개의 개체에 대하여 세 번의 이분화 분지를 반복한 의사결정나무의 사례이며[8], 트리를 구성하는 각 노드(node)를 분지하는 기준은 지니 지수(Gini index) 또는 엔트로피 지수(Entropy index)와 같은 불순도 값이 최소화하도록 결정된다.



<그림 1> 의사결정나무 사례[10]

신경망 모형은 인간의 신경망을 흉내 낸 기계학습(machine learning) 기법이다[11]. 신경망 모형은 <그림 2>의 사례와 같이 입력층, 은닉층, 출력층과 각 층(layer)에 속하는 노드들로써 구성된다. 은닉층은 1개 이상으로 구성될 수 있고, 입력층 및 출력층을 구성하는 노드의 수는 입력 및 출력의 속성 수 또는 속성의 표현방식에 따라 결정될 수 있다. 신경망 모형에

서 입력층의 노드들에 입력특성의 값이 주어지면, 하위 노드와 상위 노드를 연결한 각 화살표에 부여된 가중치를 통해 노드로의 가중합으로 계산된 값은 활성함수를 통해 상위 노드의 출력값이 된다. 활성함수로는 시그모이드(sigmoid) 함수, 소프트맥스(softmax) 함수, 렐루(ReLU) 함수 등이 있으며, 본 논문에서는 연산속도가 빠르고 최근 신경망 분야에서 가장 빈번히 활용되는 렐루 함수를 사용하기로 한다. 렐루 함수의 활성함수는 가중합 x에 대해 max(0, x)의 값을 출력한다.



<그림 2> 신경망 모형 사례

2.2 신경망 모형의 입력특성 선택을 위한 의사결정나무 활용

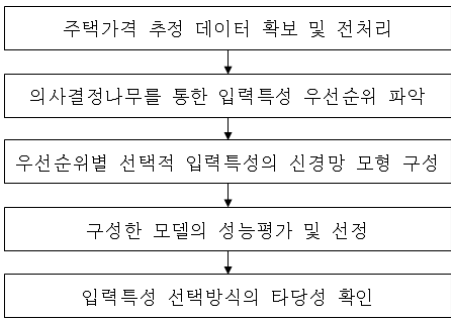
신경망 모형에서 입력특성 또는 입력변수의 선택을 위한 의사결정나무의 활용은 신경망 모형과 의사결정나무의 통합모형으로 불리기도 한다[6, 7]. 의사결정나무를 통해 신경망 모형의 입력특성을 선택하는 방식은 크게 세 가지이다[7]. 첫 번째는 구성된 의사결정나무의 분지기준에 적용된 설명변수들을 신경망 모형의 입력특성(feature)으로 선택하는 방식인데[6], <그림 1>의 경우 최상위 계층인 (100, 40)의 데이터를 (54, 40) 및 (46, 0)으로 분류한 기준을 비롯하여 (54, 40) 및 (14, 40)의 분류에 각각 사용된 설명변수들을 신경망 모형의 입력특성으로 활용할 수 있다. 두 번째는 의사결정나무의 말단노드(terminal node)에 해당되는 범주(category)의 분류규칙(if-then 형식)별로 각각의 조건절(if절)에 따라 새로이 정의되는 범주변수를 해당 신경망 모델의 입력특성으로 활용하는 경우[9]이며, <그림 1>의 경우 4개의 말단노드에 해당하는 분류규칙에 따라 범주변수를 1, 2, 3, 4 등으로 설정하여 활용할 수 있다. 세 번째는 의사결정나무의 분지에 사용된 설명변수와 말단노드의 범주변수를 한꺼번에 신경망 모델의 입력특성으로 활용하는 경우[7, 8]이며, 이상의 두 방식을 결합한 형태로 볼 수 있다.

의사결정나무로부터 신경망 모형의 입력특성을 선택하는 이상의 세 방식에 대해 신경망 모형의 성능(정확도, 재현율 등)을 고려하는 입장에서 다음과 같이 평가할 수 있다. 첫 번째로 의사결정나무 분지에 사용된 설명변수들을 입력특성(feature)으로 선택하는 경우, 불순도를 크게 낮추는 변수를 선택할 가능성이 크므로 신경망 모형의 성능에 긍정적일 가능성이 크다. 그렇지만 구성된 의사결정나무의 깊이(depth) 수가 작아서 유효한 변수가 배제되거나, 또는 의사결정나무의 깊이 수가 커서 신경망 모형의 성능에 적절하지 않은 변수까지 선택되거나 과적합(overfitting)의 가능성이 있다. 두 번째로 의사결정나무의 분류 범주를 신경망 모델의 입력특성으로 활용하는 경우에는 개체 수가 적고 불순도가 낮은 범주가 초래하는 과적합의 우려가 있고, 비교적 다양한 속성의 개체들이 동일 범주로 묶이는 경우 신경망 모형의 성능에 부정적인 효과가 있을 수 있다. 세 번째인 의사결정나무의 분지변수와 범주변수를 함께 신경망 모델의 입력특성으로 활용하는 경우에는 앞선 두 가지 방식의 문제점을 같이 가질 수 있으며, 분지변수와 범주변수의 값들이 가지는 정보가 서로 중복적인 면이 있으므로 범주변수에 의한 신경망 모델의 성능개선 효과가 기대에 미치지 않을 수 있다.

Ⅲ. 연구 범위 및 내용

신경망 모형의 입력특성 선택을 위해 의사결정나무의 분지에 적용된 변수를 활용하는 경우, 두 가지 대안을 생각해볼 수 있다. 첫 번째는 말단노드의 불순도가 0일 때까지 의사결정나무의 분지를 진행하고, 분지에 활용된 모든 변수를 신경망 모형의 입력특성으로 선택하는 경우이다. 이 경우, 앞서 살펴본 바와 같이 과적합에 의한 신경망 모형의 성능손실을 우려할 수 있다. 두 번째로는 의사결정나무의 분류성능을 최대화하는 깊이로 분지 또는 가지치기(pruning)하고, 이에 대응하는 분지변수를 활용하는 것이다. 이 경우에 정보이득(information gain)을 최대화하는 방향으로 불순도를 낮추는 분지기준의 변수로써 의사결정나무를 분지하지만, 전체 최적(global optimum)이 보장되지는 않는다. 또한 신경망 모형의 계산방식과 다르므로, 의사결정나무에서 구한 분류성능이 신경망 모형의 성능으로 반드시 이어지지 않는다. 이러한 문제점의 개선을 위해 본 논문에서는 신경망 모형의 최적 성능을 위한 입력특성의 선택방안을 제시하고, 사례 데이터에 적용하기로 한다.

이와 같은 연구범위에 대하여 본 논문에서 구성한 연구내용은 <그림 3>과 같이 단계별로 구성하였다. 첫 번째 단계에서는 주택가격 추정에 활용할 모형의 입력특성으로서 가치가 있는 데이터를 확인 및 수집, 그리고 필요한 전처리를 수행한다. 본 논문에서 활용할 모형은 의사결정나무와 신경망 모형이므로, 무의미한(dummy) 데이터의 제거나 데이터값의 정규화 등을 처리하는 것이 필요하다. 두 번째 단계에서는 확보한 데이터로써 의사결정나무를 구성하고, 깊이(depth)가 얕고 정보이득이 큰 변수의 순서대로 우선순위를 정한다. 이때 말단노드의 불순도가 0이 되도록 분지하여, 모든 분지변수들에 대해 우선순위를 확인하고자 하였다. 세 번째 단계에서는 변수의 우선순위에 따라 신경망 모형의 입력특성으로 하나씩 추가하면서 훈련(training) 데이터를 통해 신경망 모델을 구성하였다. 세 번째 단계에서는 우선순위에 따라 입력특성을 추가한 여러 신경망 모형의 성능을 평가하고 평가결과에 따라 최선의 입력특성으로 구성된 신경망 모형을 선택할 수 있다. 최종 단계에서는 앞 단계의 신경망 모형 구성에 대해 타당성을 확인하기로 한다. 또한, 사례의 주택가격 데이터에 대하여 본 논문에서 정리된 방식을 적용하였다.

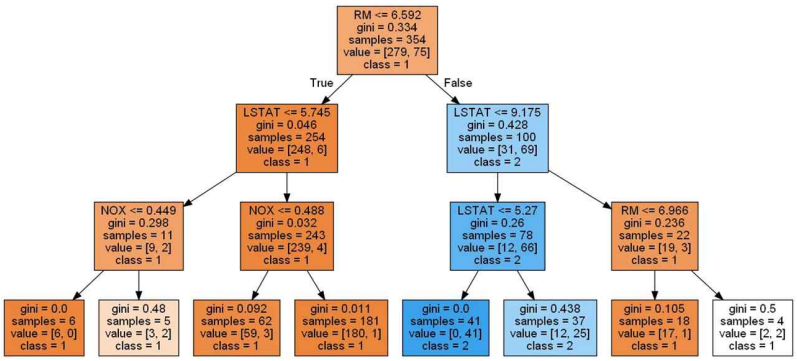


<그림 3> 연구 범위 및 내용

Ⅳ. 의사결정나무를 통한 신경망 모형의 입력특성 선택

4.1 데이터 확보 및 의사결정나무 구성

의사결정나무나 신경망 모형과 같은 데이터마이닝 기법은 지식발견을 위해 대개 다량의 데이터에 접근하지만, 목표변수에 큰 영향을 주지 않은 설명변수 또는 입력특성을 제외함으로써 모형의 성능을 개선할 수 있다[6-8]. 국가 또는 서울지역 전체와 같이 광범위한 지역의 주택가격 추정에 사용되는 대표적인 요인들은 인구구조와 경제변수 요인의 측면에서 인구수, 고령화지수, 출산률, GNP, 물가, 실업률, 가계부채 등의 데이터를 활용할 수 있다[1]. 특정 국소지역 및 특정 형태의 주택에 대해서는 용도지역, 도로



<그림 4> 구성된 의사결정나무 사례

접면, 방위, 경과연수, 건물 연면적, 토지면적, 건물구조, 지붕구조 등과 같이 개별 주택의 특성을 반영하는 데이터를 가격추정의 설명변수로 활용하기도한다[2]. 주택가격 추정을 위해 이상과 같이 분석범위나 목적, 주택의 특성, 수요공급의 특수성 등에 따라 주택가격 및 여러 설명변수를 구성할 수 있다.

주택가격 분야에서 데이터마이닝 데이터로 자주 활용되는 보스톤 주택가격 데이터[12, 13]의 변수는 <표 1>과 같으며, MEDV는 주택가격이고 나머지는 설명변수에 해당된다. 의사결정나무 분류를 적용하기 위해서는 주택가격(MEDV)을 범주형으로 변환할 수 있다. <표 1>에서 상위 3개의 설명변수와 등간격의 두 등급(하위, 상위)으로 구분한 주택가격의 데이터로 구성한 의사결정나무는 <그림 4>와 같다.

<표 1> 주택가격 등급분류를 위한 데이터 사례

변수명	설명
RM	주택 1가구당 평균 방의 개수
LSTAT	모집단의 하위계층의 비율(%)
NOX	10ppm 당 농축 일산화질소
CRIM	자치시(town) 별 1인당 범죄율
ZN	25,000평방피트를 초과하는 거주지역 비율
INDUS	비소매상업지역이 점유하고 있는 토지의 비율
AGE	1940년 이전에 건축된 소유주택의 비율
DIS	5개의 보스턴 직업센터까지의 접근성 지수
RAD	방사형 도로까지의 접근성 지수
TAX	10,000 달러 당 재산세율
PTRATIO	자치시(town)별 학생/교사 비율
B	$1000(Bk-0.63)^2$ (Bk: 자치시별 흑인의 비율)
MEDV	본인 소유의 주택가격(중앙값) (단위: $1,000)

(참조: https://rpubs.com/chocka314/251613)

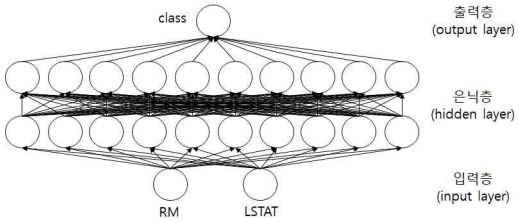
4.2 입력특성 선택 및 신경망 모형 구성

의사결정나무의 분지는 전체 불순도를 낮추도록 정보이득이 큰 변수로써 상위 계층(layer)에서 먼저 이루어지며, 또 분지된 그룹의 불순도가 작을수록 재분지된 결과의 정보이득이 크고 불순도가 작을 가능성이 크다. 따라서 구성된 의사결정나무로부터 다음 규칙의 순서대로 신경망 모형에 선택할 입력특성의 우선순위를 정하기로 한다.

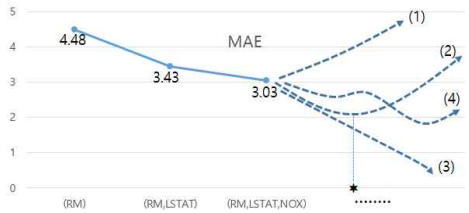
(1) 상위 계층의 분지에 활용된 설명변수가 우선
(2) 동일 계층에 복수의 분지변수가 있으면, 불순도가 낮은 그룹의 분지변수가 우선
(3) 앞 순위에 이미 선택된 분지변수는 무시
(4) 우선순위로 입력특성을 1개씩 추가해가며 신경망 모형의 입력특성을 구성

위 규칙에 따라 <그림 4>의 의사결정나무에서 구한 입력특성의 우선순위는 RM→LSTAT→NOX인데, 이 우선순위에 따라 신경망 모형의 입력변수를 선정하기로 한다. 우선순위로 입력특성을 1개씩 추가하여 신경망 모형의 입력특성을 구성하면 {(RM), (RM,

LSTAT), (RM, LSTAT, NOX)}의 세 가지이다. 신경망 모형에서 2개의 은닉층을 가지고 각 은닉층의 노드 수가 입력층 노드 수의 5배로 구성된다면, 입력변수가 (RM, LSTAT)인 경우 신경망 모형은 <그림 5>와 같다. (RM) 또는 (RM, LSTAT, NOX)에 대해서도 같은 방식으로 신경망 모형을 구성할 수 있다.



<그림 5> (RM, LSTAT)의 입력특성을 가진 신경망 모형



<그림 6> 입력특성에 따른 신경망 모형의 성능(MAE)

이상에서 구한 신경망 모형의 세 가지 입력특성에 대해 각각 신경망 모형을 구성하고, 표준화한 전체 506건의 데이터를 훈련용과 평가용으로 각각 70% 및 30%로 나누어 모형을 학습하고 가장 일반적인 성능 측정치인 평균절대오차(MAE: mean absolute error)를 사용하였다. MAE는 $(\sum_i |y_i - \tilde{y}_i|)/n$으로 계산되며, y_i는 i번째 평가 데이터의 실제 값이고 $\tilde{y}_i$는 신경망 모형에 의한 i번째 추정값이다. 신경망 모형의 구조는 실험을 통해 높은 성능을 보이는 형태로 구성하였는데, 입력층과 출력층 그리고 2개의 은닉층으로 구성하고 각 은닉층 노드는 입력층 노드 수의 5배로 구성하였다.

훈련용 데이터에 의한 250회(epochs)의 훈련으로 3가지 입력특성으로 구성한 각 신경망 모형의 MAE의 변화추이는 <그림 6>과 같다. (RM)→(RM, LSTAT)→(RM, LSTAT, NOX)의 입력특성으로 구한 신경망 모형의 순서로 MAE가 작으므로, 순서대로 모형의 성능이 더 우수해진다고 볼 수 있다.

4.3 입력특성이 신경망 모형의 성능에 미치는 영향

<표 1>의 전체 12개 설명변수(NOX, … , B) 모두에 대해 순서대로 구성한 신경망 모형의 MAE는 <그림 6>의 (1), (2), (3), (4)의 점선과 같이 변화할 수 있다. (1)과 (2)의 경우는 각각 (RM, LSTAT, NOX) 또는 가장 우수한 성능을 보이는 ✶에 해당하는 입력특성과 같이, 모형의 성능에 가장 유리한 입력특성을 선택하는 경우이다. 모형의 성능에 가장 유리한 입력특성 선택 전후의 신경망 모형에서, 입력특성 우선순위에 따라 MAE가 감소 및 증가하는 모습을 보인다. 이는 의사결정나무의 불순도 감소의 방향으로 선택되는 분지변수가 신경망 모형의 성능에도 영향을 미치는 것으로 이해할 수 있다.

그리고 <그림 6>의 (3)에서는 <표 1>의 모든 설명변수가 입력특성으로서 가치가 있는 것으로 볼 수 있다. 신경망 모형의 성능변화가 다소 불규칙적인 (4)의 경우는 전체 설명변수를 입력특성으로 선택한 경우보다 우수하거나, 또는 최저 MAE를 보이는 모형을 최종 모형으로 선택할 수 있다. 신경망 모형의 경우, 모형의 구성이나 학습과정 등에 따라 모형의 성능이 다소 불규칙적인 면이 <그림 6>의 (4)와 같은 형태를 보일 수 있는 것으로 판단할 수 있다.

Ⅴ. 사례 데이터를 통한 적용 및 평가

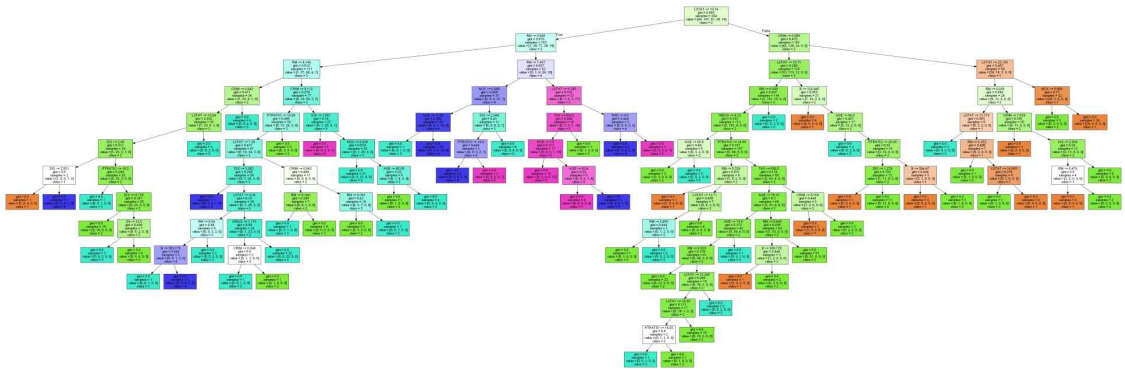
5.1 사례 데이터 및 의사결정나무 구성

정리한 방식에 따라, <표 1>의 변수를 포함하는

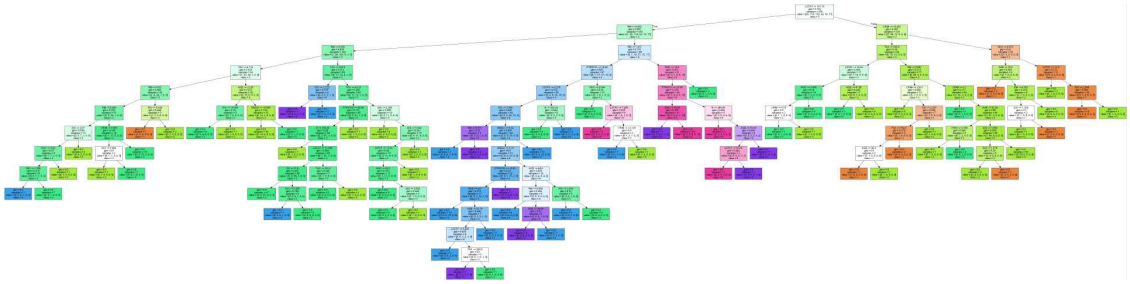
506건의 보스톤 주택가격 데이터[12, 13]에 대해 분석하기로 한다. 주택가격을 등간격으로 5등급, 6등급, 7등급으로 구분한 각각에 대해 <그림 7~9>와 같이 말단노드의 불순도가 0이도록 의사결정나무를 구성하였다. 그리고 구성한 의사결정나무로부터 신경망 모형의 입력특성으로 선택할 설명변수의 우선순위를 앞서 정한 규칙에 따라 구하면 <표 2>와 같다. 7등급 의사결정나무의 마지막 순위인 INDUS는 분지변수로 선택되지 않은 경우이다.

5.2 입력특성 우선순위에 따른 신경망 모델

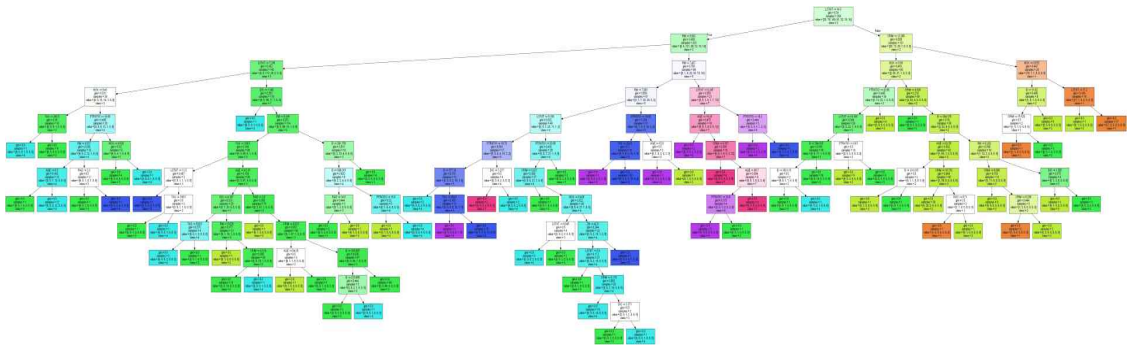
<표 2>에서 정리된 입력특성의 우선순위에 따라 <그림 7~9>의 의사결정나무 각각에 대해 12개씩의



<그림 7> 5등급의 주택가격으로 분류한 의사결정나무



<그림 8> 6등급의 주택가격으로 분류한 의사결정나무



<그림 9> 7등급의 주택가격으로 분류한 의사결정나무

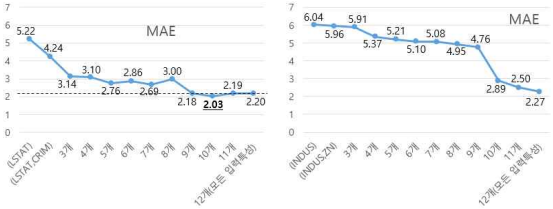
신경망 모형을 구성하고 성능을 평가하였다. 12개의 신경망 모형 각각은 <표 2>의 입력특성 순위에 따라 입력특성이 1개로부터 12개까지 구성되며, 각 신경망 모형에 대해 평가 데이터로써 구한 MAE는 <그림 10>과 같다. <그림 10>의 (1)~(3)에서 좌측 그림은 <표 2>의 우선순위에 따라 1개의 입력특성(LSTAT)을 가진 신경망 모형으로부터 12개 입력특성 모두를 가진 신경망 모형까지 MAE를 나타내고 있다. 그리고 우측 그림은 우선순위의 역순으로 입력특성을 하나씩 추가하여 구성한 신경망 모델의 MAE를 나타낸다.

<표 2> 주택가격 등급 수 및 입력특성 순위(지니지수,깊이)

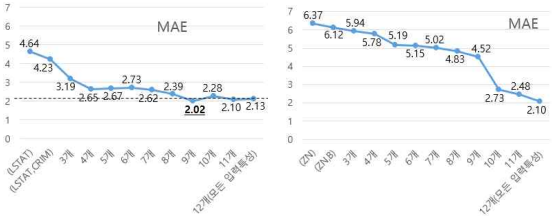
순위	5등급	6등급	7등급
1	LSTAT (0.683, 1)	LSTAT (0.723, 1)	LSTAT (0.74, 1)
2	CRIM (0.475, 2)	CRIM (0.483, 2)	CRIM (0.565, 2)
3	RM (0.679, 2)	RM (0.665, 2)	RM (0.659, 2)
4	NOX (0.17, 4)	TAX (0.34, 3)	NOX (0.447, 3)
5	DIS (0.263, 4)	NOX (0.42, 3)	DIS (0.263, 4)
6	B (0.553, 4)	RAD (0.431, 4)	B (0.406, 4)
7	AGE (0.095, 5)	PTRATIO (0.607, 4)	PTRATIO (0.495, 4)
8	TAX (0.204, 5)	DIS (0.095, 5)	TAX (0.18, 5)
9	RAD (0.444, 5)	AGE (0.245, 5)	AGE (0.475, 5)
10	PTRATIO (0.466, 5)	INDUS (0.08, 6)	RAD (0.5, 7)
11	ZN (0.408, 9)	B (0.09, 7)	ZN (0.244, 9)
12	INDUS (0.08, 9)	ZN (0.035, 9)	INDUS (-, -)

<그림 10>의 좌측 그림에서 점선은 모든 입력특성을 포함한 신경망 모형의 MAE인데, 최소 MAE는 (1), (2), (3)순서대로 10번째, 9번째, 9번째의 우선순위 입력특성까지 포함한 신경망 모형이다. 즉, 사례 데이터에서 2~3개의 입력특성(INDUS, ZN 등)은 신경망 모형의 성능에 좋지 않다는 것을 의미하며, INDUS 및 ZN 등을 제외한 입력특성의 신경망 모형이 가장 우수한 모형으로 선정될 수 있다. 또한 <그림 10>에서 좌측 그림의 점선은 의사결정나무의 모든 분지변수를 신경망 모형의 입력특성으로 구성하는 기존의 방식에서 보이는 성능수준이라고 볼 수 있는데, 위로부터 각각 2.03<2.20, 2.02<2.13, 1.98<2.19와 같이 최선의 입력특성을 선택함으로써 신경망 모형의 성능이 개선됨을 볼 수 있다.

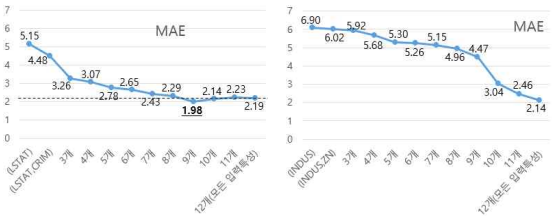
<그림 10>의 (1), (2), (3)에서 우측 그림은 우선순위의 역순으로 입력특성을 하나씩 추가하여 구성한 신경망 모델의 MAE를 나타내는데, 모든 등급(5, 6, 7)에서 좌측의 우선순위로 구한 신경망 모형보다 대체로 성능(MAE)이 좋지 않다. 이는 INDUS, ZN 등을 비롯한 우선순위가 낮은 입력특성이 신경망 모형의 성능에 좋지 않은 영향을 미친다는 것을 의미한다.



(1) 5등급의 주택가격 의사결정나무로 선택한 입력특성



(2) 6등급의 주택가격 의사결정나무로 선택한 입력특성



(3) 7등급의 주택가격 의사결정나무로 선택한 입력특성

<그림 10> 의사결정나무로부터 선택한 입력특성의 신경망 모형 성능(MAE)

5.3 입력특성 선택방식의 타당성

사례의 주택가격 데이터를 통해, 본 논문의 우선순위에 의한 신경망 모형의 입력특성 선택방식은 다음

측면에서 타당함을 보인다.

(1) 한 신경망 모형의 성능을 <그림 10>에서 비교할 때, 모든 입력특성이 포함되는 마지막 순위를 제외하고는 동일한 수의 입력특성에서 우선순위에 의한 신경망 모형이 역순의 경우에 비해 모형성능이 거의 모두 우수하다. 또한, 성능향상에 있어서 우선순위에 의한 경우는 입력특성의 수가 늘어날수록 성능향상이 초기에 가파르다가 점차 완만하게 되며, 역순의 경우에는 그 반대로 나타난다. 이는 의사결정나무로써 정한 입력특성 우선순위가 신경망 모형의 성능에 유리하게 작용한다는 것을 의미한다.

(2) 주어진 데이터로부터 신경망 모형에 가능한 입력특성이 n개인 경우, 입력특성의 선택가능한 대안의 수는 $\sum_{i=1}^{n} {}_nC_i$이므로 최적의 선택이 쉽지 않다. 이런 경우, 본 논문의 방식을 통해 적절한 수준의 성능을 가지는 신경망 모형을 구성할 수 있다.

VI. 결론 및 토의

최근에 활용이 확대되고 적용방식이 다양해지는 빅데이터 분야[14, 15]는 한층 유연하고 우수한 성능의 분석방식을 요구한다고 볼 수 있다. 이러한 측면에서, 본 논문에서는 빅데이터 또는 데이터마이닝 분야의 대표적인 분석방식인 신경망 모형에서 입력특성 선택의 우선순위를 위해 의사결정나무를 활용하는 방식을 정리하였다.

이는 신경망 모형 및 의사결정나무의 보완적인 또는 결합적인 분석방식이라고 할 수 있으며, 의사결정나무의 분지변수를 신경망 모형의 입력특성으로 활용하는 기존의 방식과 달리 신경망 모형의 성능을 최대화하는 입력특성을 우선순위에 따라 선별적으로 선택하는 방식이라고 할 수 있다. 제안한 방식으로 주택가격 추정에 적용하여 타당성을 확인한 결과, 주어진 입력특성을 모두 활용하는 일반적인 방식보다 우선순위에 따른 적절한 입력특성의 선택은 모형의 추정력을 높일 수 있다는 점을 여러 비교를 통해 정리하였다.

본 논문에서 정리한 방식은 주택가격 추정문제 이외에도 신경망 모형을 활용하는 다른 영역에서 입력특성의 선택에 활용이 가능할 것으로 생각된다. 이를 통해 의사결정나무 및 신경망 모형을 활용하는 여러 분야에서 분석성능의 개선과 활용범위의 확대에 도움이 될 것으로 기대할 수 있다.

본 논문의 분석방식 및 적용사례는 다소 실험적인 측면에서 통해 타당성의 확인이 이루어진 면이 크다. 따라서 본 논문에서 다룬 신경망 모형의 입력특성과 의사결정나무를 통한 우선순위에 대해 이론적이고 수치적인 측면의 추가적인 분석이 필요한 것으로 생각된다. 한편, 본 논문의 분석방식에 대해 다양한 적용과 경험적 타당성이 확보된다면 실용적인 데이터 분석방식으로 활용될 것이라 기대된다.

참고문헌

[1] 최정일 · 이옥동, “디지털 경제에 부동산 가격의 변동에 영향을 주는 요인에 관한 연구,” 디지털정책연구, 제11권, 제11호, 2013, pp.59-70.

[2] 이창로 · 박기호, “단독주택가격 추정을 위한 기계학습 모형의 응용,” 대한지리학회지, 제51권, 제2호, 2016, pp.219-233.

[3] 이지영 · 유재필, “인공신경망을 이용한 주택가격지수 예측,” 한국산학기술학회논문지, 제22권, 제4호, 2021, pp.228-234.

[4] Adetunjia, A.B. et al., “House Price Prediction Using Random Forest Machine Learning Technique,” Procedia Computer Science, Vol.199, 2022, pp.806-813.

[5] Kang, J. et al., “Developing A Forecasting Model for Real Estate Auction Prices Using Artificial Intelligence,” Sustainability, Vol.12, No.7, 2020, 2899.

[6] 강진웅 · 금기정 · 손승녀, “의사결정나무와 신경망 모형 결합에 의한 운전자 우회결정요인 분석,” 한국도로학회논문집, 제13권, 제3호, 2011, pp.167-176.

[7] 이극노 · 이홍철, “이동통신고객 분류를 위한 의사결정나무(C4.5)와 신경망 결합 알고리즘에 관한 연구,” 한국지능정보시스템학회논문지, 제9권, 제1호, 2003, pp.139-155.

[8] 서광규 · 안범준, “하이브리드 의사결정나무와 인공신경망 모델을 이용한 방문학습지사의 고객세분화,” 한국산학기술학회논문지, 제7권, 제3호, 2006, pp.518-523.

[9] 김광섭 · 박정아, “의사결정나무 분류와 인공신경망을 이용한 토양수분 산정모형 개발,” 대한토목학회논문집, 제31권, 제2B호, 2011, pp.155-163.

[10] 최병호 · 조남욱, “머신러닝을 이용한 선불전자지급수단의 이상금융거래 탐지 연구,” 한국전자거래학회지, 제27권, 제2호, 2022, pp.65-77.

[11] McCulloch, W.S. and Pitts, W., “A Logical Calculus of The Ideas Immanent in Nervous Activity,” The Bulletin of Mathematical Biophysics, Vol.5, No.4, 1943, pp.115-133.

[12] Adetunjia, A.B. et al., “House Price Prediction Using Random Forest Machine Learning Technique,” Procedia Computer Science, Vol.199, 2022, pp.806-813.

[13] Li, D.K., Mei, C.L. and Wang, N., “Tests for spatial dependence and heterogeneity in spatially autoregressive varying coefficient models with application to Boston house price analysis,” Regional Science and Urban Economics, Vol.79, 2019, 103470.

[14] 정병호, “빅데이터 분류 기법에 따른 벤처 기업의 성장 단계별 차이 분석,” 디지털산업정보학회 논문지, 제15권, 제4호, 2019, pp.197-212.

[15] 윤한성, “속성유사도에 따른 사회연결망 서브그룹의 군집유효성,” 디지털산업정보학회 논문지, 제17권, 제1호, 2021, pp.75-84.

■ 저자소개 ■

윤 한 성
(Yoon Han-Seong)

2001년 3월~현재 경상대학교 경영대학 교수
1998년 8월 한국과학기술원 테크노경영대학원 (공학박사)
1987년 8월 한국과학기술원 산업공학과 (공학석사)
1985년 2월 서울대학교 산업공학과(공학사)

관심분야 : 디지털경영, 기술경영, 공급사슬, 데이터분석 등
E-mail : hsyun@gnu.ac.kr

논문접수일 : 2023년 2월 15일
수정접수일 : 2023년 2월 28일
게재확정일 : 2023년 3월 02일