KJCEM 15. 4. 068~076
JULY 31, 2014

www.kicem.or.kr (pISSN 2005-6095)
http://dx.doi.org/10.6106/KJCEM.2014.15.4.068

유전 알고리즘 - 서포트 벡터 회귀를 활용한 공동주택 공사비 예측에 관한 연구

남군[1] • 최재웅[2] • 최혜미[1] • 김주형*
[1]한양대학교 첨단건축도시환경공학과 · [2]삼성에버랜드

A Study on Estimating Construction Cost of Apartment Housing Projects Using Genetic Algorithm – Support Vector Regression

Nan, Jun[1] · Choi, Jae-Woong[2] · Choi, Hyemi[1] · Kim, Ju-Hyung*
[1]Department of Frontier Architectural and Urban Environmental Engineering, Hanyang University
[2]Samsung Everland

Abstract : The accurate estimation of construction cost is important to a successful development in construction projects. In previous studies, the construction cost are estimated by statistical methods. Among the statistical methods, support vector regression (SVR) has attracted a lot of attentions because of the generalization ability in the field of cost estimation. However, despite the simplicity of the parameter to be adjusted, it is not easy to find optimal parameters. Therefore, to build an effective SVR model, SVR's parameters must be set properly without additional data handling loads. So this study proposes a novel approach, known as genetic algorithm (GA), which searches SVR's optimal parameters, then adopt the parameters to the SVR model for estimating cost in the early stage of apartment housing projects. The aim of this study is to propose a GA-SVR model and examine the feasibility in cost estimation by comparing with multiple regression analysis (MRA). The experimental results demonstrate the estimating performance based on the percentage of estimations within 25% and find it can effectively do the accurate estimation without through the trial and error process.

Keywords : Cost estimate, Genetic algorithm, Support vector regression

1. 서론

1.1 연구의 배경 및 목적

1.1.1 연구의 배경

일반적으로 건축프로젝트 초기 기획단계에서 설계가 완료된 후, 총 공사비용의 80%가 결정되기 때문에, 조정 가능한 비용이 20% 미만이다(Duverlie 1999). 소요예산이 결정되면 유효범위 내에서 공법 및 대안을 선정하기 때문에 프로젝트 초기 단계에서의 정확한 공사비 예측이 개발사업의 성패를 좌우한다고 할 수 있다.

* Corresponding author: Kim, Ju-Hyung, Department of Architectural Engineering, Hanyang University, Seongdong-Gu, Seoul, 133-791, Korea
E-mail: kcr97jhk@hanyang.ac.kr
Received February 10, 2014: revised April 4, 2014
accepted April 23, 2014

전통적으로 통계적 분석을 활용하여 공사비를 예측 하였으나, 1980년대부터 전문가시스템(Expert System), 인공신경망(Artificial Neural Network, ANN), 사례기반추론(Case-Based Reasoning) 등과 같은 다양한 인공지능 기법을 이용하여 공사비를 예측하는 방법들이 활발하게 연구되었다(Lee 2012). 그중에서도 인공신경망은 비용 예측분야에서 회귀분석보다 우수한 것으로 나타나면서 점점 더 활용되고 있다(Bode 2000). 그러나 인공신경망은 입력패턴 분포추정을 위한 대량의 학습 데이터가 필요하고, 과도적합(over fitting) 문제로 인해 일반화가 어려울 뿐만 아니라 초기화 작업에서 연구자가 공사비 예측에 대한 경험이 없으면 변수설정에 어려움을 겪고 결과해석이 어렵다는 한계가 있다(Park 2006). 이러한 문제의 해결방안으로 서포트 벡터 회귀(Support Vector Regression, SVR) 방법이 제시되었다. Vapnik(1995)에 의해 제안된 서포트 벡터 회귀는 결과 해석이 용이하고, 적은 학습 데이터만으로도 신

속하게 분별학습을 수행할 수 있는 일반화 능력을 가지고 있다. Kwon(2009)과 Park(2007)은 서포트 벡터 회귀 모델로 초기단계에서의 소프트웨어개발 비용이나 공동주택 공사비 비용을 예측할 경우, 사례기반 추론, 인공신경망이나 퍼지 논리보다 더 우수한 예측 성능 및 모델구축에 있어서 뛰어난 일반화 능력이 있음을 증명하였다.

그러나 서포트 벡터 회귀 모델은 조정할 파라미터가 단순함에도 불구하고 여전히 최적의 파라미터를 결정하는 과정에서 다소 시행착오적인 방법을 거쳐야 하는 문제가 있다. 상이한 파라미터 설정은 예측성능에 큰 영향을 미치기 때문에 최적화된 파라미터를 선택하는 것은 서포트 벡터 회귀를 설계하는데 중요한 과정이다. 따라서 본 연구에서는 유전 알고리즘이라는 방법을 도입해서 기존의 시행착오적인 과정을 거치지 않고서도 높은 예측정확성을 확보할 수 있는 GA-SVR 모델을 제시하여, 제안하는 공동주택의 공사비 예측모델의 타당성과 실효성을 자체평가와, 다른 기법과의 비교를 통해서 검증하는 것을 연구의 목적으로 한다.

1.2 연구의 범위 및 방법

본 연구는 공동주택 프로젝트 초기 기획단계에 공사비 예측을 위한 GA-SVR 모델을 제안한다. 서포트 벡터 회귀에서는 성능과 밀접한 관련이 있는 몇몇의 파라미터 값을 사용자의 정의에 의존하게 되는데, 파라미터 값에 따른 성능 변화를 예측하기 어렵다. 따라서 사용자 정의 파라미터의 최적 값을 구하기 위한 방법으로 유전 알고리즘을 사용한다. 개발언어로는 파이슨(Python)을 사용하며, 개발 알고리즘은 개방된 라이브러리(open source library)를 기반으로 LIBSVM(a LIBrary for Support Vector Machines)를 활용하여 유전 알고리즘과 서포트 벡터 회귀의 결합 모델인 GA-SVR을 구축한다.

수집된 실제사례 데이터들을 가지고 교차검증(Cross Validation)을 통해서 GA-SVR 결합 모델의 성능을 측정하고, 기존 연구에서 사용한 방법들과 비교하여 평가한다.

본 연구의 절차는 다음과 같다.

공사비 예측방법 및 서포트 벡터 회귀와 유전 알고리즘에 관련된 기존의 연구 및 문헌을 고찰한다. GA-SVR 모델 제안하고 공동주택 프로젝트 초기 기획단계 공사비 예측에 영향을 미치는 요인을 추출한 후, 수집한 실제 공사비를 활용하여 제안한 GA-SVR 모델의 성능평가를 진행해 예측 정확성을 검증한다.

2. 공사비 예측방법에 관한 이론고찰

2.1 기존 공사비 예측방법

건축 프로젝트 초기단계에 빠르고 정확한 공사비 예측의 필요성이 인식되면서 다양한 방법론을 이용한 연구가 진행되었다. 2000년대 이후부터 공사비 예측에 회귀분석 모델, 신경망 모델, 사례기반추론을 이용한 모델, 유전 알고리즘 결합 모델, 가중치 모델, 서포트 벡터 회귀분석 예측 모델과 공사비의 민감도 분석 등 여러 가지 많은 방법론들이 제시되었음을 알 수 있다(Kim 2013). 최근에 Jin(2014)은 사례기반추론으로 사업초기단계 공사비 및 공사기간 예측모델, 표준 S-curve 예측모델, 발주자 관점을 고려한 현금흐름 예측모델을 제안하고 타당성과 실효성을 검증하였다. Cho(2013)는 학교시설 공사비 예측에 관한 연구에서 전통적인 분석방법인 회귀분석방법과 인공신경망 네트워크방법을 적용해서, 두 가지 방법을 비교하여 인공신경망을 적용했을 경우 평균오차율과 표준 분포측면에서 더 우수한 결과를 보이면서 뛰어난 공사비 예측성능을 보임을 증명했다. Son(2012)은 주성분 분석방법 (Principal Component Analysis, PCA)을 서포트 벡터 회귀에 적용해서 공사비를 예측할 수 있는 PCA-SVR 모델을 개발하였다. 연구에서 기존에 진행되었던 다중선형회귀 (Multiple Linear Regression), 의사결정나무 (Decision Tree, DT), 인공신경망, 서포트 벡터 회귀를 단독으로 사용 할 때 보다 PCA-SVR모델이 예측정확도가 높다는 것을 검증했다. 서포트 벡터 회귀는 인공신경망보다 높은 예측력을 나타낼 뿐만 아니라 인공신경망의 한계점으로 지적되었던 과대적합으로 인한 일반화 어려움, 국소최적화와 같은 문제점들을 완화하는 장점을 갖고 있기 때문에 많은 관심을 받고 있다(Park 2009).

2.2 서포트 벡터 회귀와 파라미터

서포트 벡터 머신(Support Vector Machine)은 1979년 Vapnik에 의해 제안된 통계적 학습이론으로 초평면(hyperplane)을 발견하여 두 범주를 갖는 객체들을 분류하는 방법이다. 서포트 벡터 머신을 회귀문제에 적용하여 학습 데이터에 의존한 서포트 벡터 회귀 예측모델을 만들 수 있다. 서포트 벡터 회귀 모델은 커널 함수를 통하여 입력 데이터에서 해결할 수 없는 비선형 회귀문제를 고차원 공간으로 사상(mapping)시켜 선형 회귀문제로 전환한 후 해결할 수 있다. 그리고 손실함수를 도입하여 오차범위를 조절함으로써 과대적합을 피면하여 일반화 성능을 높이는 장점이 있어 학습과정을 통하여 좋은 예측을 할 수 있다(Vapnik 1995).

Kwon(2009)은 서포트 벡터 회귀 모델을 소프트웨어개발 비용 예측에 활용했고, Park(2007)은 초기단계 공동주택 공사비 예측에 활용했으며, Chen(2007)은 관광수요 부문에 서포트 벡터 회귀를 적용하여 관광수요를 예측했다.

$$f(x) = \overline{w} \bullet \overline{x} + b \quad (1)$$

$$\overline{w} = \sum_{i=1}^{n} \alpha_i y_i \overline{x}_i \quad (2)$$

$$b = -\sum_{i=1}^{n} \alpha_i y_i \overline{x}_i \bullet \overline{x}_j \quad (3)$$

$$f(a_i, a_i^*) = \sum_{i=1}^{n} f(x_i)(a_i - a_i^*) - \epsilon \sum_{i=1}^{n} (a_i + a_i^*) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (a_i - a_i^*)(a_j - a_j^*) k(x_i, x_j) \quad (4)$$

$$K(x, x_i) = \exp\left(-\gamma \parallel X - X_i \parallel^2\right), \text{ for } \gamma > 0 \quad (5)$$

$$L_\epsilon(y) = \begin{cases} 0 & \text{for} |f(x) - y < \epsilon| \\ |f(x) - y| - \epsilon & otherwise \end{cases} \quad (6)$$

$$\alpha_i, \alpha_i^* \in [0, C] \quad (7)$$

일반적으로 서포트 벡터 회귀에서 구하는 초평면을 식(1)로 표시하는데, 식에서 나타나는 상수 w와 b는 식(2)와 식(3)에서처럼 라그랑지 승수 αi, αi*로 표시할 수 있다. 가지고 있는 데이터와 식(5), 식(6), 식(7)을 통하여 C, γ, ε를 확정함으로써, αi, αi*를 구할 수 있다. 실제 사용되는 학습 데이터는 오류 데이터를 포함하고 있는 경우가 일반적이다. 오류 데이터로 인하여 선형분리가 불가능할 경우 가능해가 존재하지 않게 된다. 이러한 경우에 올바른 분리경계면을 기대하기 위해서는 오류 데이터를 허용하는 것이 필요한데, 상수 C는 추정오차의 패널티로 모델의 일반화 성능과 복잡도(Model complexity) 를 결정하는 모수이다. 비선형 커널함수의 파라미터 γ는 입력 데이터의 표현 공간에서 초평면을 결정하는 비선형을 구축하는데 사용되는데 식(5)는 RBF (Radial Basis Function) 의 커널함수 이다. ε은 서포트 벡터 회귀 모델에 사용되는 손실함수로 회귀모델의 경우 목표 값과 예측 값 사이의 오차가 발생하게 되는데 이러한 예측오차의 임계치를 설정하여 오차의 허용범위를 크게 해서 예측모델의 일반화 성능을 높이는데 사용된다(Park, 2006).

서포트 벡터 회귀 모델의 일반화 성능(예측정확도)과 효율성은 파라미터 C, γ, ε와 관계된다. 때문에 본 연구에서의 주요 문제는 학습 데이터를 통해 각각의 매개 변수에 대한 별도의 최적화 된 파라미터를 찾는 것이 아니라 세 변수의 상관관계를 고려한 최적의 파라미터 셋을 찾는 것이다.

기존의 최적 파라미터를 선정한 연구를 고찰하면, Scholkopf(1999), Cherkassky(1998)와 Vapnik(1995)등 많은 연구자들이 기존의 경험과 전문지식을 바탕으로 C와 ε를 선택했으나, 이런 접근방법들은 비전문가 사용자에게는 적합하지 않는 것으로 판단된다. Scholkopf(2002)은 최적 서포트 벡터 회귀 파라미터 세팅을 찾기 위하여 그리드 검색 최적화 방법(Grid search optimization method)을 제안했으나 이 방법을 활용하기엔 시간의 소비가 많음을 알 수 있었다. Pai(2005)는 GA-SVR 모델을 개발하여 전기회로 예측방법을 제안했으며, 유전 알고리즘의 구현을 이진 코드 유전 알고리즘을 사용하여 파라미터를 결정하였으나 예측의 정확성이 부족하였다. 이런 문제점들을 감안하여 Chen(2007)은 관광수요 예측을 목적으로 학습 데이터로 부터 모든 파라미터를 동시에 최적화하는 GA-SVR 모델을 제안했으며, 실험을 통해 관광수요 예측분야에서 GA-SVR 모델은 전형적인 신뢰성 예측도구로 예측성능이 뛰어나다는 것을 검증했다.

2.3 Genetic Algorithm

유전 알고리즘은 진화의 원리를 문제 해결에 이용하는 대표적인 방법론 중 하나로써 Holland(1975)에 의해서 제안되었다. 교차와 돌연변이에 의한 변화로 환경에 더욱 잘 적응 할 수 있는 새로운 개체를 만들어 내는 진화 과정에서 많은 시행착오들이 발생한다. 그러나 이러한 과정을 인식하고 유전 알고리즘을 적용한 통계학적 분석방법으로 많은 문제들을 해결할 수 있다. Kim(2004)은 공동주택 공사비 예측을 위해 신경회로망에 유전 알고리즘을 적용했으며, 기존의 신경회로망으로 초기단계 공사비를 예측할 때보다 유전 알고리즘을 적용 후, 효율적이고 정확하게 공사비가 예측됨을 검증했다. Park(2007)은 서포트 벡터 회귀 모델을 적용하여 초기단계 공동주택의 공사비를 예측했으며, 기존의 신경망 모델과 비교하여 예측 정확성과 모델 구축의 용이성이 우수하고 건축 프로젝트의 예산 산정과 조정에 유용하게 활용될 수 있음을 밝혔다. 그러나 서포트 벡터 회귀 모델은 조정할 파라미터가 단순하지만 최적의 파라미터를 결정하는 방법에 있어서 다소 시행착오적인 한계가 있다고 밝혔다. 때문에 본 연구에서는 최적의 파라미터를 결정하기 위해 유전 알고리즘을 적용하고, 도출된 최적 파라미터가 서포트 벡터 회귀 모델을 거쳐 공사비 예측을 할 수 있는 GA-SVR 모델을 제안하려고 한다.

3. GA-SVR 모델 제안

기존의 연구들에서 사용한 방법들을 보면, 파라미터를 선택할 때 많은 시행착오적인 방법들을 사용하였다. 먼저 부동한 파라미터 셋을 기반으로 서포트 벡터 회귀 모델을 만든 다음 그 모델들을 각각 하나씩 평가하여 최적화된 파라미터를 결정하였다. 그러나 이런 과정은 많은 시간과 운이 필요하다(Chen 2007).

위에서 말하는 절차와는 달리, 본 논문에서는 GA-SVR라는 모든 파라미터를 동시에 최적화 하는 방법을 제안하려고 한다. Chen(2007)의 기본 모형을 토대로 서포트 벡터

회귀모형을 구축하는데 필요한 여러 개의 파라미터들 중에서, 공사비 예측에 활용할 수 있는 파라미터를 선정해서 Fig. 1과 같이 GA-SVR 모델을 제안한다.

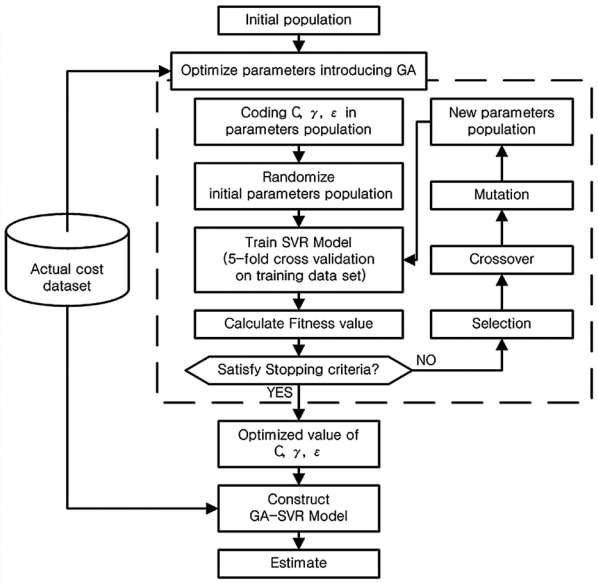

Fig. 1. Proposed GA-SVR model

염색체로 구성된 초기 개체군은 유전 알고리즘을 거쳐 임의로 생성된다. 세 파라미터(i.e., C, γ, ε)의 값은 직접 염색체에 코딩된다. 제안한 모델의 구체적인 내용은 아래와 같다.

(1) 염색체 표현: 이진 코드로 변환 할 필요 없이 직접 염색체를 파라미터 실수 값으로 구성할 수 있다. 유전자 X는 X={p1, p2, p3}으로 표현되고 여기서 p1, p2, p3은 각각 정규화한 파라미터 C, γ, ε를 대표한다.

(2) 적합도 정의: 생태계에서 개체들은 환경에 대한 적응여부에 따라 적자생존의 원칙으로 자연적으로 도태되고 진화한다. 이러한 환경에 적응한다는 개념을 도입하여 정량화해서 적합도 라고 한다. 본 연구에서는 최고의 예측정확도를 가지고 있는 예측모델의 최적화 된 파라미터를 찾기 위하여, 데이터를 학습시켜 구축된 모델의 예측정확도를 적합도로 한다. 따라서 예측정확도는 평균 제곱근 오차(Root Mean Square Error, RMSE)를 통해 계산한다.

평균 제곱근 오차는 예측한 값과 실제로 관찰하여 얻은 값을 비교할 때 사용하는 측도이다. 평균 제곱근 오차 값이

$$Minf = RMSE_{cross\ validation} \quad (8)$$

$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\alpha_i - e_i)^2} \quad (9)$$

α_i = 실제 값
e_i = 예측 값
n = 학습데이터 샘플수

0에 가까울수록 최적화된 파라미터 값이라는 것을 알 수 있다(식 8, 식 9).

(3) 개체군 초기화: 본 연구에서의 초기 개체군은 15개의 임의로 생성된 염색체들로 이루어진다.

(4) 적합도 평가: 각각의 염색체 적합도 평가는 식(8)과 식(9)에 근거하여 계산된다.

(5) 선택: 대표적인 방법으로 룰렛 휠(Roulette Wheel)을 사용하여 염색체를 선택한다.

(6) 교차: 교차 연산자는 simulated binary crossover을 사용한다. simulated binary crossover은 실수 벡터에 대해서도 이진 문자열에 대한 1점 교차 연산과 같은 효과를 얻을 수 있도록 고안된 교차 연산자이다. 새로운 염색체를 생성하는 확률을 0.8로 한다.

(7) 변이: 변이는 교차 후에 발생한다. 제안하는 모델에서는 다음 세대에서의 변이 발생 여부를 polynomial mutation method에 따라서 결정한다. 변이 발생 확률을 0.05로 설정한다.

4. GA-SVR 모델 검증

4.1 데이터 수집 및 구축

공동주택 공사비를 예측할 때, 관련성이 낮은 요인을 변수로 입력하면 공사비 예측정확도가 떨어지기 때문에 초기 단계에서 공사비에 영향을 미치는 요인을 선택해야 할 필요가 있다. 따라서 이전 연구를 고찰함으로써 공사비 예측을 위한 요인을 선택하였다.

Lim(2010)은 공사비에 영향을 미치거나 예측하는데 필요한 모든 영향요인을 감안할 경우, 공사비 예측의 정확도는 향상되겠지만 수집할 수 있는 데이터의 한계가 발생하고 데이터를 수집하는데 소요되는 인력 및 비용의 낭비가 발생할 수 있을 뿐만 아니라 이후에 공사비 예측방법이 복잡해짐에 따라 투입되는 시간이 증가하여 효율성이 떨어질 수 있다고 판단하고, 도출된 영향요인에 대한 전문가 면담을 실시하여 중복되는 영향요인을 통합하고, 공동주택의 특성을 반영할 수 없는 요인을 제거하였다. 결론적으로 공동주택 초기공사비에 영향을 미치는 영향요인을 연면적, 용적률, 건폐율, 대지면적, 건축면적, 층수, 지하층수, 동수, 총 세대수, 평균평수, 마감수준, 층고, 지붕형식, 주차장면적, 기초형식 등 15개 요인으로 설정하였다.

건설 사업은 프로젝트의 성격에 따라 다른 특성을 갖기 때문에, 본 연구에서는 공동주택을 중심으로 한 수도권, 전남 지역의 33개의 케이스를 바탕으로 초기 기획단계에서 데이터를 수집할 수 있는 요인을 감안하여 기초형식, 지붕형식, 마감수준, 연면적, 대지면적, 건축면적, 용적률, 건폐율, 평균층수, 지하층수, 동수 와 총세대수를 변수로 설정

Table 1. Data for cost estimation

	Foundation types	Roof types	Finishing grades	Gross floor area	Lot area	Building area	Building coverage ratio (%)	Floor area ratio (%)	Average Story	Basement story	Building numbers	Total house numbers	Actual cost (Thousand Won)
Case1	PHC+MAT	RCSR	2	37,427	24,550	4,266	17.38	116.70	12	2	7	282	29,100,914
Case2	PHC+MAT	RCSR	3	65,319	34,097	5,589	16.39	152.03	14	2	9	442	45,781,135
Case3	PHC+MAT	RCSR	3	112,934	46,915	8,313	17.72	173.30	15	2	13	987	76,025,436
Case4	PHC+MAT	RCSR	3	92,428	38,390	6,637	17.29	179.01	15	2	10	748	58,215,038
Case5	PHC+MAT	RCSR	4	84,759	36,913	5,876	15.92	170.55	15	2	9	722	56,663,196
Case6	PHC+MAT	RCSR	4	64,077	29,879	4,945	16.55	164.25	15	2	10	553	46,929,199
Case7	PHC+MAT	RCSR	4	80,750	32,357	5,800	17.92	179.18	15	2	11	731	53,104,295
Case8	PHC+MAT	RCSR	3	52,879	27,454	5,165	18.81	144.81	14	2	9	410	39,236,237
Case9	PHC+MAT	RCSR	4	110,664	52,189	8,081	15.48	163.37	15	2	13	841	76,452,682
Case10	PHC+MAT	RCSR	3	94,038	40,972	6,984	17.05	177.57	15	2	12	694	67,139,437
Case11	PHC+MAT	RCSR	3	145,885	49,092	10,152	20.68	223.82	15	2	16	1,421	88,168,262
Case12	PHC+MAT	RCSR	4	78,498	26,733	5,694	21.30	228.28	15	2	9	629	48,920,046
Case13	PHC+MAT	RCSR	4	75,332	25,183	5,650	22.44	228.59	15	2	8	518	46,115,640
Case14	PHC+MAT	RCSR	3	108,554	36,283	7,437	20.50	226.99	15	2	14	919	63,401,352
Case15	PHC+MAT	RCSR	4	96,761	36,582	8,658	23.67	193.74	15	2	19	948	61,028,217
Case16	PHC+MAT	RCSR	3	46,020	17,311	4,000	23.11	196.57	13	2	6	399	25,870,068
Case17	PHC+MAT	RCSR	3	71,766	27,798	6,314	22.71	194.89	13	2	9	643	42,939,946
Case18	PHC+MAT	RCSR	4	17,514	12,655	3,730	29.47	98.79	4	1	7	115	12,945,721
Case19	PHC+MAT	LSSR	1	43,736	8,592	2,163	25.18	378.79	20	2	5	319	21,697,629
Case20	MAT	LSSR	1	158,458	30,360	6,806	22.42	380.72	23	2	12	1,060	73,450,105
Case21	PHC+MAT	RCSR	2	33,897	12,264	2,109	17.20	200.45	21	1	2	260	16,999,513
Case22	PHC_MAT	RCSR	1	114,848	48,849	10,893	22.30	228.60	15	2	8	591	54,138,342
Case23	PHC+MAT	RCSR	2	74,924	28,544	7,096	24.86	237.08	14	3	7	392	44,768,787
Case24	PHC+MAT	RCSR	3	49,106	15,538	2,899	18.66	189.79	15	2	5	237	24,132,033
Case25	PHC+MAT	RCSR	3	37,123	11,501	2,120	18.44	219.39	14	2	4	183	19,022,160
Case26	PHC+MAT	RCSR	3	116,558	40,227	7,614	18.93	169.45	18	1	10	577	74,311,636
Case27	MAT	RCSR	4	44,210	20,068	3,937	19.62	239.32	12	1	5	216	31,401,950
Case28	PHC+MAT	RCSR	4	46,476	13,825	2,535	18.34	189.59	13	1	4	200	34,239,244
Case29	PHC+MAT	RCSR	5	77,958	14,305	7,698	53.82	421.00	23	3	4	315	62,790,969
Case30	PHC+MAT	RCSR	3	69,020	26,742	4,201	15.71	208.93	17	1	7	443	51,769,203
Case31	PHC+MAT	RCSR	4	213,976	68,262	16,550	24.25	238.93	15	1	17	1,281	138,613,933
Case32	MAT	RCSR	3	93,057	30,100	4,937	16.44	237.43	20	1	11	553	58,660,234
Case33	PHC+MAT	RCSR	3	41,662	29,536	2,829	9.58	102.61	15	1	4	204	16,305,748

하였다(Table 1).

수집된 데이터는 입력 변수들의 특성 및 크기가 다르기 때문에 이를 표준화 하였다. 기초형식은 PHC파일 사용 유무에 따라 1과 0으로 입력하였다. 지붕 형식의 경우 철근콘크리트 위 슁글 지붕(Reinforced Concrete capped by Shingles Roof, RCSR)을 이용한 사례에 1을, 경량철골 경사 지붕(Light Steel Sloping Roof, LSSR)을 사용한 사례에 0을 입력 하였다. 수량적 요인에 해당하는 공사비 영향요인 에서는 각각의 영향요인 별로 그 단위가 다르게 나타났다. 따라서 모든 영향요인을 0에서 1사이의 수로 표준화 하여 사용하였다. 표준화 방법은 식 10 과 같다.

$$A_{SV} = \frac{A_{OV}}{\max(A_1, A_2 \ldots A_N)} \quad (10)$$

$A_{SV} = Standardized\ value\ of\ A_{ov}$
$A_{ov} = Original\ value$
$A_1 \ldots A_N = All\ original\ value$
$N = The\ number\ of\ the\ case$

그리고 자재비 및 인건비는 매년 물가상승률에 따라 그 가격이 상승하며, 공사 수행시기마다 특정 자재의 원자재 값 폭등과 같은 현상이 발생하여 시기별로 제반비용이 달라질 수 있으므로 건설공사비 지수(Korea Institute of Construction Technology 2013)를 이용하여 2010년 1월 기준으로 공사 수행시기를 보정하였다. 또 지역별로 자재비 및 노무비의 차이가 발생하고, 대도시에서 떨어진 지역일수록 자재의 운송비가 상승하기 때문에 지역별 차이가 발생한다. 이에 지역별 공사비 지수(Korean National Housing Corporation, 2005)를 이용하여 수도권을 기준으로 공사비를 보정하였다(Table 2).

4.2 예측정확도 평가기준 및 방법

보정한 데이터를 케이스 1부터 18, 케이스 19부터 33으로, 2개의 데이터 셋으로 나누어서 GA-SVR 모델의 성능평가를 수행한다. 케이스 1부터 18을 가지고 모델의 자체 성능평가를 통해 모델의 타당성을 검증하고, 케이스 19부터 33으로 다른 예측방법과 비교하여 제안하는 모델의 실효성을 검증한다. 모델 자체의 예측 정확도를 측정하기

Table 2. Results of data standardization and correction

	Foundation types	Roof types	Finishing grades	Gross floor area	Lot area	Building area	Building coverage ratio (%)	Floor area ratio (%)	Average Story	Basement story	Building numbers	Total house numbers	Actual cost (Thousand Won)
Case1	1	1	0.4	0.175	0.360	0.258	0.323	0.277	0.522	0.667	0.368	0.198	0.21144
Case2	1	1	0.6	0.305	0.499	0.338	0.305	0.361	0.609	0.667	0.474	0.311	0.33263
Case3	1	1	0.6	0.528	0.687	0.502	0.329	0.412	0.652	0.667	0.684	0.695	0.55237
Case4	1	1	0.6	0.432	0.562	0.401	0.321	0.425	0.652	0.667	0.526	0.526	0.42297
Case5	1	1	0.8	0.396	0.541	0.355	0.296	0.405	0.652	0.667	0.474	0.508	0.41169
Case6	1	1	0.8	0.299	0.438	0.299	0.308	0.390	0.652	0.667	0.526	0.389	0.34097
Case7	1	1	0.8	0.377	0.474	0.350	0.333	0.426	0.652	0.667	0.579	0.514	0.38584
Case8	1	1	0.6	0.247	0.402	0.312	0.349	0.344	0.609	0.667	0.474	0.289	0.28508
Case9	1	1	0.8	0.517	0.765	0.488	0.288	0.388	0.652	0.667	0.684	0.592	0.55548
Case10	1	1	0.6	0.439	0.600	0.422	0.317	0.422	0.652	0.667	0.632	0.488	0.48781
Case11	1	1	0.6	0.682	0.719	0.613	0.384	0.532	0.652	0.667	0.842	1.000	0.64060
Case12	1	1	0.8	0.367	0.392	0.344	0.396	0.542	0.652	0.667	0.474	0.443	0.35543
Case13	1	1	0.8	0.352	0.369	0.341	0.417	0.543	0.652	0.667	0.421	0.365	0.33506
Case14	1	1	0.6	0.507	0.532	0.449	0.381	0.539	0.652	0.667	0.737	0.647	0.46065
Case15	1	1	0.8	0.452	0.536	0.523	0.440	0.460	0.652	0.667	1.000	0.667	0.44341
Case16	1	1	0.6	0.215	0.254	0.242	0.429	0.467	0.565	0.667	0.316	0.281	0.18796
Case17	1	1	0.6	0.335	0.407	0.381	0.422	0.463	0.565	0.667	0.474	0.452	0.31199
Case18	1	1	0.6	0.1947	0.4327	0.1710	0.1780	0.2437	0.6522	0.3333	0.2353	0.1593	0.09406
Case19	1	0	0.2	0.204	0.126	0.131	0.468	0.900	0.870	0.667	0.263	0.224	0.20379
Case20	0	0	0.2	0.741	0.445	0.411	0.417	0.904	1.000	0.667	0.632	0.746	0.65201
Case21	1	1	0.4	0.158	0.180	0.127	0.320	0.476	0.913	0.333	0.105	0.183	0.14879
Case22	1	1	0.2	0.537	0.716	0.658	0.414	0.543	0.652	0.667	0.421	0.416	0.47077
Case23	1	1	0.4	0.350	0.418	0.429	0.462	0.563	0.609	1.000	0.368	0.276	0.37957
Case24	1	1	0.6	0.229	0.228	0.175	0.347	0.451	0.652	0.667	0.263	0.167	0.20538
Case25	1	1	0.6	0.173	0.168	0.128	0.343	0.521	0.609	0.667	0.211	0.129	0.16189
Case26	1	1	0.6	0.545	0.589	0.460	0.352	0.402	0.783	0.333	0.526	0.406	0.52852
Case27	0	1	0.8	0.207	0.294	0.238	0.365	0.568	0.522	0.333	0.263	0.152	0.22467
Case28	1	1	0.8	0.217	0.203	0.153	0.341	0.450	0.565	0.333	0.211	0.141	0.24556
Case29	1	1	1	0.364	0.210	0.465	1.000	1.000	1.000	1.000	0.211	0.222	0.44711
Case30	1	1	0.6	0.323	0.392	0.254	0.292	0.496	0.739	0.333	0.368	0.312	0.36691
Case31	1	1	0.8	1.000	1.000	1.000	0.451	0.568	0.652	0.333	0.895	0.901	1.00000
Case32	0	1	0.6	0.435	0.441	0.298	0.305	0.564	0.870	0.333	0.579	0.389	0.41587
Case33	1	1	0.6	0.195	0.433	0.171	0.178	0.244	0.652	0.333	0.211	0.144	0.11672

위해서 리브 원 아웃 교차검증 (leave-one-out cross-validation) 방법을 사용한다. 이 방법은 초기 n개의 데이터 셋을 n-1개의 학습 데이터와 1개의 검증 데이터로 분할한 후, 각각의 데이터를 대상으로 검증을 n번 반복 수행한다.

예측 정확도는 평균 제곱근 오차와 PRED(25) (percentage of PREDictions within 25%)를 사용해서 오차율(error)로 평가한다. PRED(25)는 예측 값이 실제 값의 25%범위 내에 있는 비율을 나타내며, 좋은 모델일수록 PRED(25)의 수치는 높게 나타나게 된다. 사례기반추론은 도출된 결과를 이해하기 쉽고, 신규사례를 데이터베이스에 저장하는 것 이외에 추가적인 작업 없이도 학습을 진행할 수 있다는 장점이 있어서 경험 지향적인 문제해결 기법으로 공사비 예측에 가장 적합하다(Jin 2014). 때문에 마지막으로 사례기반 추론방법으로 예측하여 얻은 결과와 비교하여 제안한 GA-SVR 모델의 실효성을 평가한다.

4.3 예측정확도 분석 및 평가

실제 공동주택 공사비 데이터를 이용한 파라미터의 2500세대까지의 진화과정을 보면(Fig. 2), 평균 제곱근 오차 값이 가장 0에 가까운 세대는 1824번째 세대로 나타났으며 이때의 평균 제곱근 오차 값은 0.006976이고 파라미터 값은 각각 C=180.5986, ε=0.001003, γ=0.001251 로 나타났다. 이후 5000세대까지 측정한 결과 평균 제곱근 오차가 0.006844로 0.0001의 차이만을 보였다. 따라서 2500세대까지의 파라미터 값 중 최적 값을 선택하여 공사비 예측을 실시하였다.

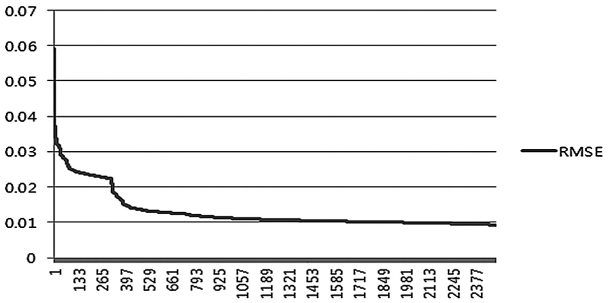

Fig. 2. RMSE trends through data analysis

도출된 파라미터 값을 적용하여 구한 예측 공사비를 실제 공사비와 비교해 모델의 타당성을 검증했다(Table 3,

Fig. 3). 케이스 16과 케이스 18의 경우 오차율이 이상적으로 크게 나와 데이터 특성이 정확히 반영되지 않은 것으로 판단되어 평균오차율의 계산에서 제외하였다.

케이스 19부터 케이스 33을 대상으로 GA-SVR 모델의 타당성을 검증하고(Table 4), 동일한 데이터로 사례기반 추론방법과 비교하여 제안하는 모델의 실효성을 검증했다.

Table 3. GA-SVR model cost estimation results(Case1-18)

Case	Actual cost	Estimated cost	Error percentage	PRED(25)
Case1	29,813,582	31,966,027	7.22%	O
Case2	46,902,289	46,748,311	0.33%	O
Case3	77,887,343	75,963,044	2.47%	O
Case4	59,640,701	60,593,192	1.60%	O
Case5	58,050,867	56,971,190	1.86%	O
Case6	48,078,554	47,816,567	0.54%	O
Case7	54,404,754	56,811,290	4.42%	O
Case8	40,197,067	41,635,461	3.58%	O
Case9	78,325,023	75,047,779	4.18%	O
Case10	68,783,622	62,938,532	8.50%	O
Case11	90,327,529	96,191,373	6.49%	O
Case12	50,118,054	52,822,111	5.40%	O
Case13	47,245,072	49,372,700	4.50%	O
Case14	64,954,061	71,251,072	9.69%	O
Case15	62,522,848	76,772,837	22.79%	O
Case16	26,503,626	34,840,560	31.46%	X
Case17	43,991,518	51,947,738	18.09%	O
Case18	13,262,811	27,542,411	107.67%	X

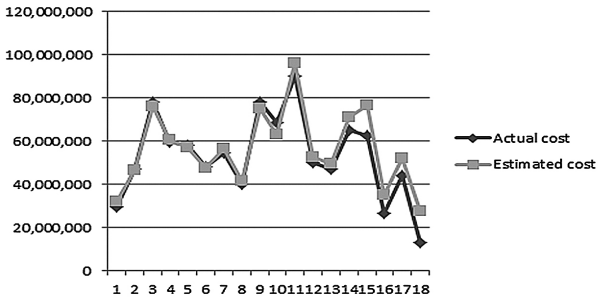

Fig. 3. Comparison of the actual cost and estimated cost

사례기반 추론방법에서의 전진제거 방식과 후진제거 방식으로 분석한 결과, 전진제거 사례기반 추론방법을 사용한 경우 평균오차율과 PRED(25)가 각각 12.07%, 70%로 나타났고, 후진제거 사례기반 추론방법을 사용한 경우에는각각 17.51%, 90%로 나타났다. 제안한 GA-SVR을 적용했을 경우 각각 5.31%, 93.33%로 우수한 예측 정확도를 보인다. 이전 연구의 경우, 상관관계 분석을 통해 서포트 벡터 회귀 모델의 파라미터를 결정하기 위해 여러 차례 선별과정을 거치는데, 본 연구에서는 이러한 과정을 거치지 않고서도 우수한 예측 정확도를 나타냄을 알 수 있었다(Table 5).

Table 4. GA-SVR model cost estimation results(Case19-33)

Case	Actual cost	Estimated cost	Error percentage	PRED(25)
Case19	28,736,161	32,746,068	13.95%	O
Case20	91,937,104	78,645,669	14.46%	O
Case21	20,980,028	24,601,324	17.26%	O
Case22	66,380,752	71,279,273	7.38%	O
Case23	53,521,497	47,122,539	11.96%	O
Case24	28,959,937	28,958,386	0.01%	O
Case25	22,827,760	24,082,284	5.50%	O
Case26	74,523,804	66,020,201	11.41%	O
Case27	31,680,210	37,342,416	17.87%	O
Case28	34,625,809	29,174,970	15.74%	O
Case29	63,044,427	52,489,339	16.74%	O
Case30	51,736,089	44,385,909	14.21%	O
Case31	141,005,233	127,083,927	9.87%	O
Case32	58,640,269	56,975,843	2.84%	O
Case33	16,457,567	30,050,189	82.59%	X

Table 5. Accuracy evaluation of GA-SVR model cost estimation

Model	Mean percentage error	PRED(25)
Forward selection method	12.07%	70%
Backward elimination method	17.51%	90%
GA-SVR	5.31%	93.3%

5. 결론

최소한의 정보를 가지고 공사비를 예측하는 모델은 프로젝트 초기 기획단계에 필요하다. 본 연구에서는 서포트 벡터 회귀를 도입하여 공사비를 예측하였다. 서포트 벡터 회귀는 분류 문제에 있어서 뛰어난 일반화 능력을 보이지만, 데이터 집합에 따라 시행착오적인 방법으로 적합한 파라미터 값을 매번 찾아야 하고, 몇 개의 특정 값을 임의로 정해서 그 값이 최적화 된 값인지에 대한 객관성이 부족한 단점이 있다. 이러한 단점을 보완하기 위해 서포트 벡터 회귀에서 사용되는 파라미터 값을 유전 알고리즘으로 최적화해서, 그 적용가능성을 검토하였다. 본 연구에서 제안한 GA-SVR 모델은 유전 알고리즘을 통해 세대가 진화하면서 최적 파라미터를 얻었기 때문에 시행착오를 거치지 않으면서도 최적화 된 파라미터 값을 찾을 수 있었다. 제안한 모델의 자체평가를 통해서 객관적으로 최적의 파라미터를 찾아가는 과정을 보여주었고, 실제 값과 예측 값을 비교한 결과 기존 연구에서 제안한 사례기반 추론방법보다 더 좋은 예측정확도를 나타내었다. 최적 파라미터를 찾는 과정을 단순화 시키면서 정확도를 향상시켰기 때문에 서포트 벡터 회귀를 적용한 연구 및 공사비 예측 실효성을 높일 수 있을 것으로 기대된다.

이번 연구에서는 공동주택의 공사비 예측으로 범위를 한정했다. 건축 프로젝트는 유형에 따른 다양한 특성이 존재하고, 이는 공사비 예측 방법에 영향을 미칠 수 있기 때문

에 향후 공동주택 이외의 다른 유형의 프로젝트에 적용해 서포트 벡터 회귀 공사비 예측 방법의 범용성을 검증할 필요가 있다.

References

Bode, J. (2000). "Neural networks for cost estimation: Simulations and pilot application." International Journal of Production Research, 38(6), pp. 1231-1254.

Chen, K. Y., and Wang, C. H. (2007). "Support vector regression with genetic algorithms in forecasting tourism demand." Tourism Management, 28(1), pp. 215-226.

Cherkassky, V., and Mulier, F. (1998). Learning from data: Concepts, theory, and methods, Wiley, New York, USA.

Cho, H. G., Kim, K. G., Kim, J. Y., and Kim, G. H. (2013). "A comparison of construction cost estimation using multiple regression analysis and neural network in elementary school project." Journal of the Korea Institute of Building Construction, 13(1), pp. 66-74.

Duverlie, P., and Castelain, J. M. (1999). "Cost estimation during design step: Parametric method versus case based reasoning method." The International Journal of Advanced Manufacturing Technology, 15(12), pp. 895-906.

Holland, J. H. (1975). Adaptation in natural and artificial systems, MA: MIT Press, London, UK.

Jin, R. Z. (2014). "Development of cash flow forecasting model in the early stage of construction project." Ph.D. Dissertation, The graduate School of the University of Seoul, Seoul, Korea.

Kim, G. H., Yoon, J. E., An, S. H., Cho, H. H., and Kang, K. I. (2004). "Neural network model incorporating a genetic algorithm in estimating construction costs." Building and Environment, 39(11), pp. 1333-1340.

Kim, M. J., Moon, H. S. and Kang, L. S. (2013). "Development of an approximate cost estimating model for bridge construction project using CBR method." Korean Journal of Construction Engineering and Management, KICEM, 14(3), pp. 42-52.

Korea Institute of Construction Technology (2013) "Construction Cost Index", Korea Statistical Information Service.

Korea National Housing Corporation (2005) "Public housing construction cost analysis", Korea Land and Housing Corporation, pp. 368-369.

Kwon, K. T., and Park, S. K. (2009). "Estimation of software project effort with genetic algorithm and support vector regression." The Korea Information Processing Society Transactions. Part D., 16D(5), pp. 729-736.

Lee, H. S., Lee, H. K., Park, M. S., Kim, S. Y. and Ahn, J. S. (2012). "Conceptual cost estimating system development for public apartment projects." Korean Journal of Construction Engineering and Management, KICEM, 13(4), pp. 152-163.

Lim, S. Y. (2010). "A study on improving the estimation accuracy of apartment project cost." MS thesis, Chonnam National University, Kwang Ju, Korea.

Pai, P. F., and Hong, W. C. (2005). "Forecasting regional electric load based on recurrent support vector machines with genetic algorithms." Electric Power Systems Research, 74(3), pp. 417-425.

Park, J. S. (2006). "An empirical comparison between support vector regression and neural networks." MS thesis, Dongguk University, Seoul, Korea.

Park, S. K. (2009). "Estimation of software project effort with genetic algorithm and support." MS thesis, Kangnung National University, Won Ju, Korea.

Park, U. Y., and Kim, G. H. (2007). "Construction cost of apartment housing projects based on support vector regression at the early project stage." Journal of the Architectural Institute of Korea, 23(4), pp. 165-172.

Scholkopf, B., and Smola, A. (2002). Learning with kernels, MIT Press, London, UK.

Scholkopf, B., Burges, J., and Smola, A. (1999). Advances in kernel methods: Support vector learning, MIT Press, London, UK.

Son, H. J., Kim, C. M., and Kim, C. W. (2012). "Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables." Automation in Construction, 27, pp. 60-66.

Vapnik, V. N. (1995). The nature of statistical learning theory, Springer-Verlag, New York, USA.

요약 : 건축 프로젝트에서 초기단계에서의 정확한 공사비 예측은 성공적인 프로젝트의 중요한 요소이다. 기존의 연구에서 공사비를 예측하기 위한 방법으로 통계학적인 방법이 활용되었다. 통계학적 방법 중 서포트 벡터 회기분석은 비용예측 분야에서 뛰어난 일반화 능력으로 많은 주목을 받고 있다. 하지만 서포트 회귀분석은 조정해야 할 파라미터가 단순함에도 불구하고 최적의 파라미터를 결정하는 방법은 시행착오적인 방법을 적용해야 하는 문제점이 있었다. 따라서 최적의 파라미터를 보다 효율적으로 결정하기 위해 본 연구에서는 유전 알고리즘을 적용하고, 이를 통해 서포트 벡터 회귀를 효율적으로 활용한 공사비 예측이 가능 할 것이다. 본 연구의 목적은 유전 알고리즘과 서포트 벡터 회귀를 활용하여 공동주택의 프로젝트 초기 기획단계의 공사비 예측모델을 구축하는 것이다. 유전 알고리즘을 통해 최적의 파라미터를 찾아내고, 이를 서포트 벡터 회귀모델에 적용시켜 공사비를 예측하였다.

키워드 : 공사비 예측, 유전 알고리즘, 서포트 벡터 회귀