# 서포트벡터머신을 이용한 교육시설 초기 공사비 예측에 관한 연구

# A Study on Predicting Construction Cost of Educational Building Project at early stage Using Support Vector Machine Technique

신 재 민* 김 광 희**
Shin, Jae-Min Kim, Gwang-Hee

Abstract

The accuracy of cost estimation at an early stage in school building project is one of the critical factors for successful completion. So various of techniques are developed to predict the construction cost accurately and expeditely. Among the techniques, Support Vector Machine(SVM) has an excellent ability for generalization performance. Therefore, the purpose of this study is to construct the prediction model for construction cost of educational building project using support vector machine technique. And to verify the accuracy of prediction model for construction cost. The performance data used in this study are 217 school building project cost which have been completed from 2004 to 2007 in Gyeonggi-Do, Korea. The result shows that average error rate was 7.48% for SVM prediction model. So using SVM model on predicting construction cost of educational building project will be a considerably effective way at the early project stage.

키워드 : 서포트벡터머신, 회귀분석, 공사비 예측, 교육시설
Keywords : support vector machine, regression analysis, predicting construction cost, school building

## I. 서론

### I-1. 연구의 배경 및 목적

건설프로젝트의 초기 단계에서 공사비의 예측은 건축주에게 전체 프로젝트의 사업성 등에 조언을 하기 때문에 프로젝트의 성패를 좌우하는 매우 중요한 요소이다. 특히 교육시설 건축프로젝트 초기단계에서의 정확한 공사비 예측은 예산의 준비 및 결정을 위해 매우 중요하다. 하지만 프로젝트의 초기 단계는 도면과 시방서 등 설계도서가 완전히 확정되지 않은 상태이므로 정확하고 신속하게 공사비를 산출하는 데 어려움이 있다. 그 결과 한정된 정보를 바탕으로 프로젝트의 초기단계에서 공사비를 정확하게 예측할 수 있는 기법들이 다양하게 발전되어 왔다.

대표적인 공사비 예측 방법으로는 회귀분석을 이용한 통계적 방법과 인공지능기법을 이용한 인공신경망(Artificial Neural Networks;NN), 사례기반추론기법(Case-based Readoning;CBR), 서포트벡터머신(Support Vector Machine; SVM)등이 있다.[1] 그 중에서도 1990년대 이후에는 인공신경망을 이용하여 공사비를 예측 방법이 주로 연구되었다. 그러나 인공신경망의 경우 수량적인 변수를 이용한 예측에 강하다는 장점에도 불구하고 사용자가 결정하여야 할 요소가 많기 때문에 실제로 적합한 모델을 설계하기 어렵다는 단점이 지적되고 있다.

이를 보완하기 위하여 최근에는 Vapnik(1995)에 의해 도입된 SVM이 뛰어난 일반화 능력으로 인하

* 경기대 건축공학과 석사과정
** 경기대 플랜트·건축공학과 교수, 공학박사

1) 김광희 외, 공사단계별 공사비 영향 변수의 선택방법과 데이터 개수와의 상관관계에 관한 연구, 대한건축학회논문집(구조계), 제23권, 제4호, pp.129-137, 2007

여 패턴인식과 학습이론분야에서 많은 주목을 받고 있다[2]. 특히 SVM은 한글과 한자와 같이 부류수가 많은 언어, 데이터마이닝과 같은 대용량 분류에서 모듈러 신경망보다 우수한 성능을 나타내고 있다[3]. 또한 SVM은 이미 건설 분야에서 흙막이 공법선정 모델[4] 및 개산견적의 평가[4] 등과 같은 분야에 적용되어 왔다.

본 연구에서는 건설프로젝트 초기단계에서 얻을 수 있는 한정된 정보를 바탕으로 SVM을 이용한 교육시설의 초기 공사비 예측 모델을 구축을 목적으로 한다. 이를 바탕으로 학교시설 공사비 예측에 SVM의 적용 가능성을 확인하고자 한다.

### I-2. 연구의 범위 및 방법

본 연구에서는 건설프로젝트의 초기단계에서 정확하고 신속한 공사비 산출을 위하여 SVM기법을 도입하고 그 예측 정확도를 검증하는 것이 목적이다. SVM을 이용한 교육시설의 공사비예측모델 구축을 위하여 경기도교육청에서 발주한 초등학교, 중학교, 고등학교 시설의 실적데이터를 수집하였다. 또한 모델 구축에 필요한 변수는 실제 교육시설 프로젝트의 초기 단계에서 얻을 수 있는 변수 중에서 10개를 선정하여 적용하였다.

SVM 모델은 NeuroSolution 프로그램을 사용하여 수집된 데이터를 기반으로 하여 구축하였다. SVM 모델의 정확도를 평가하기 위해서는 SPSS 19.0 프로그램을 이용한 다중회귀모델을 적용하였다. 이를 바탕으로 SVM모델과 다중회귀모델을 이용하였을 때 두 공사비 예측 모델의 예측 정확도를 비교하였다. 본 연구의 흐름은 그림1과 같다.

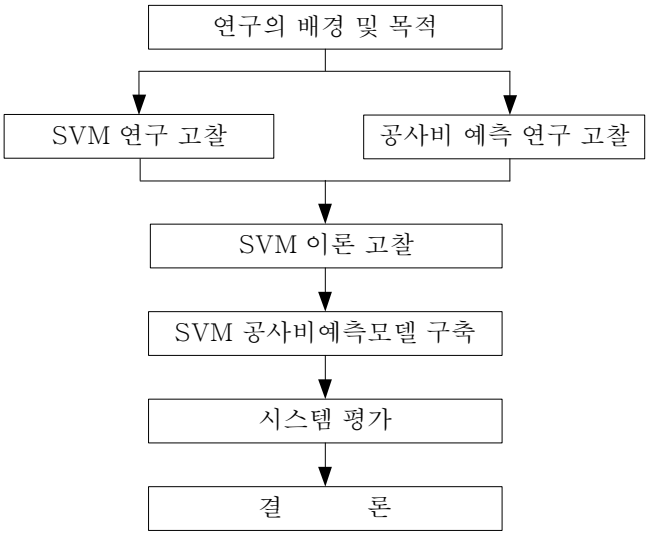


그림1. 연구의 흐름

## II. 이론적 고찰

### II-1. 선행연구 고찰

1) SVM에 관한 연구

국내 건설 산업 및 타산업에서 SVM을 이용한 기존 연구는 다음 표1과 같다.

표1. SVM에 관한 연구

| 분류 | 연구자 | 연구 내용 |
|---|---|---|
| 건설산업 | 박우열 외(2005) | 건축프로젝트 초기단계에서 실시하는 개산견적을 평가할 수 있도록 SVM을 활용한 평가모델을 구축 |
| | 박우열 외(2006) | SVM을 이용한 흙막이공법 선정모델 구축 |
| | Kim S. (2011) | SVM을 이용한 건축물 외장재료 선정모델 구축 |
| 타산업 | 김유일 외(2004) | 주가지수 예측율 향상을 위해 SVM기법을 응용하고, 주가지수 예측에서의 SVM 활용 가능성 검토 |
| | 정영미 외(2000) | SVM 학습알고리즘을 이용한 문서 범주화 실험을 수행하고, 다원 분류기로 확장할 수 있는 방안 검토 |

박우열(2005)은 개산견적을 평가하는 방법으로 SVM을 활용하여 실제 사례를 적용한 결과, 평균 77%의 높은 정확도를 보이는 것으로 나타났다. 또한 박우열(2006)과 Kim S.(2011)은 SVM을 이용하여 실제 사례를 바탕으로 높은 정확도의 흙막이공법 및 건축물 외장재료 선정모델을 구축하였다.

타 산업에서 SVM기법을 도입한 연구로 김유일

---

[2] 박우열 외, Support vector machine을 이용한 개산견적 평가모델에 관한 연구, 대한건축학회논문집(구조계), 제23권, 제4호, pp.191-198, 2005

[3] 이진선 외, 대용량 분류에서 SVM과 신경망의 성능 비교, 정보처리학회논문집 B, 제12B권, 제1호, 2005

[4] 박우열 외, Support vector machine을 이용한 흙막이 공법 선정모델에 관한 연구, 건설관리, 제7권, 제2호, pp.118-126, 2006

(2004)은 주가지수 예측에 SVM 기법을 활용하였으며, 그 결과 신경망모델보다 SVM 모델의 결과 값이 우수한 것으로 나타났다. 정영미(2000)는 SVM 학습알고리즘을 이용한 문서 범주화 실험을 SVM분류기가 높은 성능을 보임으로써 우수한 학습방법임을 증명하였다.

이를 통해 다양한 분야에서 SVM을 적용하였으며, 특히 패턴인식 및 학습이론 분야에서 높은 성능을 보인다는 사실을 알 수 있었다. 따라서 본 연구에서는 교육시설 프로젝트를 대상으로 SVM을 이용한 공사비 예측 모델을 구축하고 그 타당성을 검증하고자 한다.

2) 공사비 예측에 관한 연구

건설프로젝트의 초기단계에서 공사비를 예측할 수 있는 다양한 기법과 관련된 연구는 다음 표2와 같다.

표2. 공사비예측에 관한 연구

| 분류 | 연구자 | 연구 내용 |
|---|---|---|
| 회귀분석 | 김진원 외(2011) | 경기도 지역 BTL (Build Transfer Lease) 사업을 대상으로 회귀분석을 통해 교실의 공간비율에 따른 공사비 예측 모델 구축 |
| 인공신경망 | 김광희 외(2004) | 유전자 알고리즘을 이용하여 최적의 신경망 구조를 결정하는 방법을 공동주택의 초기 단계 공사비 예측에 적용하여 그 결과를 회귀분석 기법을 이용한 결과와 비교 |
| | 손재호 외(2008) | 회귀 분석 및 인공신경망을 이용한 BTL 교육시설물 프로젝트 공사비 예측 모델 구축 |
| 기타 | 김광희 외(2004) | 공동주택 공사비 예측분야에 사례기반추론기법을 적용하여 타당성을 검증 |

김진원(2011)은 회귀분석 기법을 이용하여 교실에 공간비율에 따른 공사비 예측 모델을 구축하였다. 이를 실제 사례에 적용한 결과 오차율은 15% 이내로 나타났다.

김광희(2004)와 손재호(2008)는 인공신경망과 회귀분석을 이용한 공사비 예측 모델을 구축하였고, 그 정확도를 비교하였다. 연구 결과 회귀분석 기법은 약 8%, 인공신경망 기법은 약 4%의 오차율을 보이며, 인공신경망 기법이 더 정확한 것으로 나타났다.

그 외의 공사비 예측 기법으로 김광희(2004)는 사례기반추론을 이용한 공동주택 공사비 예측 모델을 구축하고, 그 정확도를 회귀분석을 이용하여 검증하였다. 연구 결과로 사례기반추론은 5%의 오차율 보였으며, 회귀분석은 10%의 오차율을 보이며 사례기반추론 기법을 이용한 공사비 예측 모델이 더 정확한 것으로 나타났다.

이를 통해 현재까지 건축프로젝트 공사비 예측기법과 관련하여 진행된 연구는 회귀분석, 인공신경망 등을 이용한 연구가 대부분이었다. 또한 그 범위도 공동주택 및 교육 시설 등으로 한정적이었다. 그 중에서도 교육 시설은 BTL 사업만을 대상으로 연구가 진행되었다. 하지만 건축프로젝트의 초기단계에서 공사비 예측 방법을 활용하기 위해서는 다양한 기법 및 시설을 대상으로 한 연구가 필요할 것으로 사료된다. 따라서 본 연구에서는 초등학교, 중학교, 고등학교 교육시설을 대상으로 SVM 기법을 이용하여 공사비 예측 모델을 구축하고 그 결과를 기존의 검증된 회귀분석 기법을 이용하여 오차율을 비교, 분석하고자 한다.

## II-2. Support Vector Machine

1) SVM 개요

SVM은 1995년에 Vapnik이 제안한 통계적 학습이론으로 패턴인식과 학습이론분야에서 뛰어난 일반화 능력으로 인하여 많은 주목을 받고 있다[5]. 또한 기존의 경험적 리스크 최소화 원칙(Empirical risk minimization)이 아닌 구조적 위험 최소화(Structural risk minimization)를 기반으로 하여 일반화 오류의 상한을 최소화할 수 있다는 점에서 기존의 방법들보다 우수한 성능을 가진다고 할 수 있다[6]. 최근에는 주가지수 예측[7] 및 패턴인식문제[8], 건설 공법 선정, 공사비 예측[3] 등에 효과적으로 적용되고 있다.

SVM는 서포트 벡터 분류(Support vector

5) Vapnik V. The nature of statistical learning theory, 2nd Ed, Springer, New York, p.1-314, 1995

6) Burges CJC, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, Vol.2, No.2, pp.121-167, 1998

7) 김유일 외, 신경망과 SVM을 이용한 주가지수예측의 비교, 인터넷전자상거래연구, 제4권, 제3호, pp.221-243, 2004

8) 정영미 외, SVM 분류기를 이용한 문서 범주화 연구, 정보관리학회지, 제17권, 제4호, pp.229-248, 2000

classification;SVC)와 서포트 벡터 회귀(Support vector regression;SVR)의 두 가지 종류가 있다[9]. 그 중에서도 본 연구에서는 건설 공사비 예측을 위해 SVM의 가장 일반적인 형태인 SVR을 적용하였다. SVR은 초평면(optimal hyperplane)의 형태에 따라 선형 회귀와 비선형 회귀로 구분할 수 있다.

2) 선형 회귀 문제

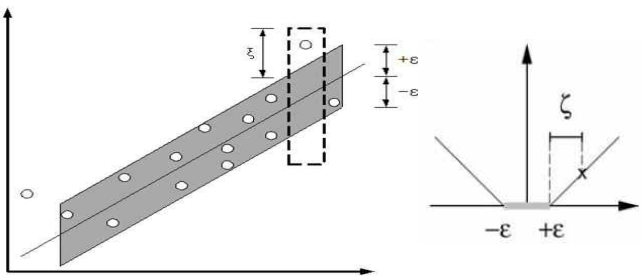


그림2. 선형회귀문제의 $\varepsilon$-무감각 손실함수

Vapnik(1995)에 의해 제안된 가장 단순한 형태의 선형 회귀 문제는 다음과 같이 정의할 수 있다[10]. 총 $l$개의 학습용 데이터가 벡터 $x_i$와 $y_i$의 쌍 $\{(x_1,y_1),(x_2,y_2),\cdots\cdots,(x_l,y_l)\}\subset \chi\times R$로 이루어져 있을 경우($\chi$는 $R^n$과 같은 입력 패턴의 공간을 나타냄), 주어진 입력 값 $x_i$에 해당하는 실제의 목표 값 $y_i$는 최고 $\varepsilon$만큼의 편차 내에 있으며, 가능한 좁은 범위의 함수 $f(x)$를 찾는 것으로 정의된다. 즉 그림2와 같이 $\varepsilon$보다 큰 편차를 허용하지 않으면서, $\varepsilon$보다 작은 오차는 무시하는 것으로 기하학적으로는 주어진 점에 적합한 가장 평평한(Flat) 초평면을 발견하는 것이라고 할 수 있다. 이때 선형함수 $f$는 식(1)과 같이 나타낼 수 있다.

$$f(x)=w \bullet x_i+b \cdots\cdots\cdots\cdots\cdots\cdots \text{식(1)}$$

식(1)에서 b는 바이어스를 말하며, 함수 $f(x)$가 가장 평평한 경우는 $w$가 가장 작은 경우이다. 이것을 푸는 방법은 유클리드 놈(Euclidean norm)인 $\|w\|^2$을 최소화하는 것으로 식(2)와 같은 볼록 최적화 문제(Convex optimization problem)로 간주할 수 있다.

$$\text{minimize } \frac{1}{2}\|w\|^2 \cdots\cdots\cdots\cdots\cdots\cdots \text{식(2)}$$

$$\text{subject to } \begin{cases} y_i-w \bullet x_i-b \le \epsilon \\ w \bullet x_i+b-y_i \le \epsilon \end{cases}$$

이 문제를 해결하기 위해서는 두 개의 슬랙변수(Slack variables) $\xi_i,\xi_i^*$를 도입하고, 이를 포함할 수 있는 새로운 최적화방정식을 유도할 수 있다. 최적화 방정식은 식(3)과 같으며 한 개는 목표 값이 $\varepsilon$이상일 경우, 그리고 다른 한 개는 목표 값이 $\varepsilon$ 이하일 경우를 위한 것이다.

$$\text{minimize } \frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*) \cdots \text{식(3)}$$

$$\text{subject to } \begin{cases} y_i-w \bullet x_i-b \le \epsilon+\xi_i \\ w \bullet x_i+b-y_i \le \epsilon+\xi_i^* \\ \xi_i,\xi_i^* \ge 0 \end{cases}$$

SVR에서는 일반화 능력을 최대화시키기 위하여 그림2와 같이 $\varepsilon$-무감각 손실함수($\varepsilon$-insensitive loss function)를 사용한다. $\varepsilon$-무감각 손실함수는 목표 값으로부터 일정거리 이내의 오차는 무시하기 때문에 슬랙변수 $\xi_i,\xi_i^*$를 도입하면 $\varepsilon$-튜브안에 있는 데이터는 무시되고, 밖에 있는 데이터의 오차가 슬랙변수 $\xi_i,\xi_i^*$로 측정된다.

식(3)에서 상수 C는 추정오차의 패널티(penalty)로서 모델의 복잡도(Complexity)를 결정하는 모수이다. C값이 크면 오차에 대해 더 큰 패널티를 할당하게 되며, 낮은 일반화 수준으로 오차를 최소화시킬 수 있다. 반면에 C값이 작으면 오차에 대해 적은 패널티를 할당하여 높은 일반화 수준을 갖게 된다. 따라서 적절한 C값의 선택은 모델의 복잡도를 결정할 뿐만 아니라 SVR의 일반화 성능도 높일 수 있다[12].

Vapnik(1995)이 제안한 최적화 문제인 식(3)을 좀 더 쉽게 해결하기 위해서는 라그랑지 승수(Lagrange multipler) $\alpha_i,\alpha_i^*,\eta_i,\eta_i^*$를 도입하며, 이는 식(4)와 같이 목적함수와 제약조건으로 구성되는 이원문제의 라그랑지 승수 함수로 나타낼 수 있다.

$$L=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)-\sum_{i=1}^{l}(\eta_i\xi_i+\eta_i^*\xi_i^*)$$
$$-\sum_{i=1}^{l}\alpha_i(\varepsilon+\xi_i-y_i+w \bullet x_i+b) \cdots\cdots\cdots \text{식(4)}$$
$$-\sum_{i=1}^{l}\alpha_i^*(\varepsilon+\xi_i^*+y_i-w \bullet x_i-b)$$

위의 식(4)에서는 쌍대 최적화 문제(Dual optimization problem)를 도출할 수 있으며, 이는 다

9) Debasish, B et al. Support Vector Regression, Neural Information Processing-Letters and Review, Vol.11, No.10, pp.203-224, 2007

10) Smola AJ et al, A tutorial on support vector regression, Neuro COLT Technical Report TR-98-030, Royal Holloway College University of London, UK, 1998

음 식(5)와 같이 표현된다.

$$\text{maximize } -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(x_i \cdot x_j) - \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \cdots \text{식(5)}$$

$$\text{subject to } \begin{cases}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C]\end{cases}$$

식(5)에서 제약조건을 만족하는 방정식을 풀게 되면 라그랑지 승수를 구할 수 있으며, 다음과 같은 SVR 추정함수를 구할 수 있다[3].

$$w=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)x_i \cdots\cdots \text{식(6)}$$

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)(x_i \cdot x)+b \cdots\cdots \text{식(7)}$$

3) 비선형 회귀 문제

선형 회귀 문제와 달리 학습용 데이터의 선형분리가 불가능할 경우를 비선형 회귀 문제라 한다. 이때는 비선형의 입력공간을 고차원의 특성 값 공간으로 사상($\Phi$, mapping) 시켜 선형 모델로 구현한 후, 다시 입력공간으로 비선형 사상을 할 수 있다[3].

이 경우 앞의 식(5)는 동일한 제약조건 하에서 식(8)과 같이 변환된다.

$$\text{maximize } -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)(\Phi(x_i) \cdot \Phi(x_j)) - \epsilon\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \cdots\cdots \text{식(8)}$$

식(8)에서 학습용 데이터 $x_i$를 비선형 사상$\Phi(x_i)$으로 변환하기 위해서는 커널함수(Kernal function) $K(x_i, x_2)=\Phi(x_1) \cdot \Phi(x_2)$를 이용하여 계산할 수 있다. 따라서 식(8)을 커널함수를 이용하여 나타내면 식(9)와 같이 표현되며, 선형문제의 식(5) 이후와 동일한 과정으로 문제를 해결할 수 있다.

$$\text{maximize } -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i \cdot x_j) - \epsilon\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)+\sum_{i=1}^{l}y_i(\alpha_i-\alpha_i^*) \text{식(9)}$$

$$\text{subject to } \begin{cases}\sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C]\end{cases}$$

## III. 공사비 예측 모델 구축

### III-1. 실적데이터 및 입력 변수

수집된 실적데이터는 경기도 교육청에서 2004년부터 2007년 사이에 준공된 초등학교, 중학교, 고등학교 교육시설 217개의 직접공사비이다. 각 교육시설 데이터의 분포는 표3과 같다.

표3. 실적데이터의 분포

| 구분(개) | 2004 | 2005 | 2006 | 2007 | 합계 |
|---|---|---|---|---|---|
| 초등학교 | 40 | 12 | 36 | 13 | 101 |
| 중 학 교 | 30 | 4 | 29 | 10 | 73 |
| 고등학교 | 11 | 7 | 14 | 11 | 43 |
| 총 합 계 | | | | | 217 |

표4. 공사비 영향 요인

| | 구분 | 최소 | 최대 | 평균 |
|---|---|---|---|---|
| 입력 변수 | 년도 | 2004년에서 2007년까지 | | |
| | 사업구분 | 1.BTL 2.재정 | | |
| | 학교구분 | 1.초등학교 2.중학교 3.고등학교 | | |
| | 용도 | 1.기존 2.택지 3.GB(Green Belt) | | |
| | 학급수 | 12 | 48 | 31 |
| | 건축면적($m^2$) | 1,204 | 3,863 | 2,694 |
| | 연면적($m^2$) | 4,925 | 12,710 | 9,656 |
| | 층 수 | 3 | 7 | 4.7 |
| | 지하층수 | 0 | 2 | 0.5 |
| | 기준층 층고(m) | 3.3 | 3.6 | 3.5 |
| 출력 변수 | 전체공사비 | | | |

본 연구에서는 건축프로젝트 초기단계에서의 공사비 예측에 필요한 SVM 모델 구축이 목적이다. 보다 정확한 공사비 예측 모델의 구축을 위해서는 공사비에 영향을 주는 요인의 개수가 많을수록 유리하겠지만, 초기단계에서의 공사비 예측에는 주어진 정보가 한정될 수밖에 없다. 따라서 공사비에 영향을 주는 요인으로 초기단계에서 수집이 가능한 정보를 바탕으로 10개를 선정하였다. 10개의 요인과 관련된 데이터의 특성은 표4와 같다.

수집된 교육시설의 실적데이터는 준공년도가 2004~2007년 사이에 분포되어 있으므로 한국건설기술연구원에서 공표하는 연도별 건설공사비지수[11]를 적용하여 자료를 보정하였다. 본 연구에서 적용

한 건설공사비 지수는 표5와 같으며, 기준년도는 2005년도이다.

표5. 연도별 건설공사비지수

| 년도 | 공사비 지수 |
|---|---|
| 2004 | 97.6 |
| 2005 | 100.0 |
| 2006 | 102.8 |
| 2007 | 106.3 |

### III-2. SVM 모델 구축

SVM 소프트웨어는 Ward Systems Group Inc.에서 개발한 NeuroSolution 6.0을 사용하였다. 수집된 총 217개의 데이터는 무작위로 20개의 테스트용 데이터(Test Data)와 37개의 교차 검증용 데이터(Cross-Validation Data), 그리고 130개의 학습용 데이터(Training Data)로 구분하였다.

구축된 SVM을 활용한 공사비 예측 모델의 표준 오차율 결과는 표6과 같다. 표6에서 평균제곱오차(MSE; Mean Squared Error)는 목표 값과 SVM모델의 예측 값 사이의 차이를 수량화한 것으로 5.55402E+11로 나타났다. 또한 선형 상관 계수(Linear Correlation Coefficient) r값은 약 0.81로 높은 상관관계를 보이는 것을 알 수 있다.

표6. 표준 오차율

| 성 과 | 전체공사비 |
|---|---|
| 평균 제곱 오차(MSE) | 5.55402E+11 |
| 선형 상관 계수 (r) | 0.811272573 |

테스트용 데이터 20개를 바탕으로 산출된 SVM 모델의 오차율은 표7과 같다. SVM 모델에 의한 예측값과 실제 프로젝트에서 집행된 목표 값과의 오차율은 식(10)을 이용하여 산출하였다.

$$\text{오차율}(\%) = \frac{|\text{목표값} - \text{예측값}|}{\text{목표값}} \times 100 \cdots\cdots \text{식}(10)$$

표7을 통해 SVM모델의 예측 공사비 평균 오차율은 7.48%로 적게 나타나는 것을 알 수 있다. 또한 4개의 데이터를 제외한 80%의 데이터가 오차율 10% 이내인 것을 통해 차이가 많이 나는 몇 개의 데이터를 제외하고는 SVM 모델의 정확도가 상당히 높은 것으로 나타났다.

표7. SVM 공사비 예측 모델의 오차율

| 오차율(%) | 비율(개수(%)) |
|---|---|
| 0-2.5 | 2 (10%) |
| 2.5-5 | 6 (30%) |
| 5-7.5 | 1 ( 5%) |
| 7.5-10 | 6 (30%) |
| 10-12.5 | 3 (15%) |
| 12.5-15 | 0 ( 0%) |
| 15-17.5 | 1 ( 5%) |
| 평균 오차율 | 7.48% |

## IV. 시스템의 평가

본 연구에서 구축한 SVM 모델을 검증하기 위해 다중회귀분석방법으로 구축한 다중회귀모델의 예측 공사비와 SVM모델의 예측 공사비를 비교하였다.

### IV-1. 다중회귀모델 구축

정확도 검증을 위하여 다중회귀모델은 SVM 모델과 동일한 실적자료를 이용하여 구축하였다.

1) 독립변수의 설정

회귀분석 모델의 독립변수는 SVM모델에서 적용하였던 변수와 동일하게 적용하였다. 독립변수 및 종속변수는 표8과 같다.

표8. 변수의 설정

| 변수 | 변수타입 | 설정 |
|---|---|---|
| 사업구분 | 명목 | BTL(X1), 재정(X2) |
| 학교구분 | 명목 | 초등학교(X3), 중학교(X4), 고등학교(X5) |
| 용도 | 명목 | 기존(X6), 택지(X7), GB(X8) |
| 학급수 | 척도 | X9 |
| 건축면적 | 척도 | X10 |
| 연면적 | 척도 | X11 |
| 층 수 | 척도 | X12 |
| 지하층수 | 척도 | X13 |
| 기준층층고 | 척도 | X14 |
| 직접공사비 | 척도 | 종속변수(Y) |

2) 회귀식의 산정

11) 한국건설기술연구원, 2012년 9월 건설공사비지수 동향, 2012

표9. 다중회귀모델

| 모델 | R | $R^2$ | 수정된$R^2$ | 추정값의 표준오차 |
|---|---|---|---|---|
| 1 | .904 | .817 | .816 | 573310.8146 |
| 2 | .923 | .852 | .851 | 516735.3232 |

a. 예측값: (상수), 연면적

b. 예측값: (상수), 연면적, 사업구분(재정)

c. 종속변수: 전체공사비

표9는 단계별 제거방식을 적용하여 산정한 다중회귀모델의 결과이다. 모델2의 결정계수($R^2$)의 값은 0.852로서 설명변수가 종속변수를 잘 설명하고 있음을 나타낸다. 다중회귀모델에 의하여 산정된 회귀식은 식(11)과 같다.

$Y= 711348.445 + 800.168X_2 - 499765.451X_{11}$ 식(11)

Y= 전체공사비 $X_2$=재정 $X_{11}$=연면적

3) 회귀식의 유효성 검증

표10은 분산분석(F검정)을 이용하여 다중회귀모델의 유의성을 검정한 결과를 보여준다. 회귀식에 의해 설명되는 분산은 3.010E14이며, 설명되지 않는 분산은 4.930E13이다. 두 평균 제곱 값의 비율인 F값은 102.696이며, 유의확률은 0.000이므로 5% 유의수준에서 통계적으로 타당함을 나타내고 있다.

회귀식에 사용되는 회귀계수의 산정과 유의수준을 검토하기 위하여 실시된 t분포에 의한 t검정 결과는 표11과 같다. t값의 유의확률에서 BTL, 건축면적, 연면적은 유의확률 5%를 초과하지 않았기 때문에 통계적으로 유의하다는 것을 알 수 있다. 하지만 건축면적과 연면적을 제외한 중학교, 고등학교, 기존, GB, 학급 수, 지상층수, 지하층수, 기준층 층고와 같은 나머지 변수들은 유의확률 5%를 초과하는 것으로 나타났다.

표10. 다중회귀모델의 분산분석(F검정)표

| 구분 | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|
| 회귀모형 | 3.010E14 | 11 | 2.737E13 | 102.696 | .000 |
| 잔차 | 4.930E13 | 185 | 2.665E11 | | |
| 합계 | 3.503E14 | 196 | | | |

## IV-2. 예측 값의 정확도 비교

본 연구에서 구축한 SVM 모델의 정확도를 평가하기 위하여 수집된 217개의 직접공사비 데이터 중에서 20개의 데이터를 무작위로 추출하여 테스트용 자료로 활용하였다. 20개의 테스트용 데이터는 회귀모델을 구축하는 자료에서 제외하였다.

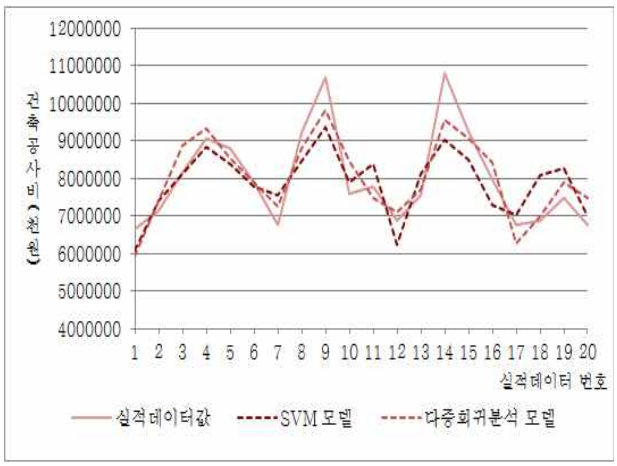


그림4. 목표값과 SVM, 회귀식의 예측 값 비교

수집된 직접공사비 데이터의 목표 값과 SVM 모델 및 다중회귀모델을 통한 공사비 예측 값의 정확도를 비교한 결과는 그림4와 표12과 같다. 비교 결과 SVM 모델의 오차율은 평균 7.48%이고, 다중회귀모델의 오차율은 평균 5.68%로 다중회귀모델의 오차율이 더 적게 나타났다.

표11. 회귀계수와 유의확률 검토

| 모델 2 | 비표준화 계수 | | 표준화 계수 | t | 유의확률 | B에 대한 95.0% 신뢰구간 | |
|---|---|---|---|---|---|---|---|
| | B | 표준오차 | 베타 | | | 하한값 | 상한값 |
| (상수) | 711348.445 | 241725.083 | | 2.943 | .004 | 234601.911 | 1188094.978 |
| 연면적 | 800.168 | 24.412 | .905 | 32.778 | .000 | 752.021 | 848.315 |
| 사업구분(재정) | -499765.451 | 73656.607 | -.187 | -6.785 | .000 | -645035.985 | -354494.916 |

표12. SVM과 회귀식의 예측 값 비교

(단위: 백만원)

| 번호 | 목표값 | SVM 모델 | | 다중 회귀 모델 | |
|---|---|---|---|---|---|
| | | 예측값 | 오차율(%) | 예측값 | 오차율(%) |
| 1 | 6,647 | 6,089 | 8.39 | 5,942 | 10.61 |
| 2 | 7,156 | 7,417 | 3.65 | 7,420 | 3.69 |
| 3 | 8,148 | 8,110 | 0.47 | 8,881 | 8.99 |
| 4 | 9,084 | 8,832 | 2.77 | 9,340 | 2.81 |
| 5 | 8,807 | 8,387 | 4.76 | 8,529 | 3.15 |
| 6 | 7,879 | 7,765 | 1.44 | 7,884 | 0.07 |
| 7 | 6,771 | 7,554 | 11.56 | 7,256 | 7.17 |
| 8 | 9,209 | 8,457 | 8.16 | 8,809 | 4.34 |
| 9 | 10,691 | 9,363 | 12.42 | 9,803 | 8.30 |
| 10 | 7,605 | 7,907, | 3.96 | 8,482 | 11.52 |
| 11 | 7,765 | 8,385 | 7.99 | 7,491 | 3.53 |
| 12 | 6,860 | 6,224 | 9.27 | 7,092 | 3.39 |
| 13 | 7,572 | 8,125 | 7.31 | 7,711 | 1.84 |
| 14 | 10,802 | 9,023 | 16.47 | 9,576 | 11.35 |
| 15 | 9,259 | 8,486 | 8.35 | 9,052 | 2.24 |
| 16 | 7,966 | 7,301 | 8.36 | 8,407 | 5.52 |
| 17 | 6,777 | 7,013 | 3.49 | 6,261 | 7.62 |
| 18 | 6,895 | 8,088 | 17.30 | 6,996 | 1.46 |
| 19 | 7,479 | 8,269 | 10.57 | 7,906 | 5.71 |
| 20 | 6,780 | 6,972 | 2.83 | 7,476 | 10.26 |
| 평균 | | | 7.48 | | 5.68 |

표13. SVM과 회귀식의 오차율 분포

| 오차율(%) | SVM | 회귀 |
|---|---|---|
| 0-2.5 | 2 (10%) | 4 (20%) |
| 2.5-5 | 6 (30%) | 5 (25%) |
| 5-7.5 | 1 ( 5%) | 3 (15%) |
| 7.5-10 | 6 (30%) | 2 (10%) |
| 10-12.5 | 3 (15%) | 3 (15%) |
| 12.5-15 | 0 ( 0%) | 0 ( 0%) |
| 15-17.5 | 1 ( 5%) | 0 ( 0%) |
| 평균 | 7.48% | 5.68% |

표14 t-test 결과

| | 검정값 = 0 | | | | | |
|---|---|---|---|---|---|---|
| | t | 자유도 | 유의확률 (양쪽) | 평균차 | 차이의 95% 신뢰구간 | |
| | | | | | 하한 | 상한 |
| 회귀 | 7.176 | 19 | .000 | 5.81123 | 4.1162 | 7.5062 |
| SVM | 7.163 | 19 | .000 | 7.47600 | 5.2914 | 9.6606 |

SVM 모델과 다중회귀모델의 오차율 분포는 다음 표13과 같다. 두 공사비 예측 모델의 오차율 분포를 보면 SVM 모델에서는 특정 사례에 대해서만 오차율이 크게 나타났다. 또한 20개의 테스트용 데이터의 오차율에 대한 t-test 결과는 표14와 같다. 표14를 통해 다중회귀모델과 SVM 모델 간의 유의 확률은 0으로 유의하지 않는 것으로 나타났다. 따라서 교육시설의 초기 공사비 예측에는 SVM모델보다 다중회귀분석 모델의 적용이 더 적합할 것으로 사료된다.

## V. 결론

정확하고 신속한 공사비의 예측은 건축프로젝트의 성패에 큰 영향을 미치는 요인 중 하나이다. 본 연구에서는 교육시설 건축 프로젝트의 초기 단계에서 공사비를 예측하는 방법으로 서포트벡터머신 기법을 이용하여 모델을 구축하였으며, 그 적용가능성을 확인하였다.

서포트벡터머신 기법을 교육시설의 공사비 예측에 적용한 결과, 상대적으로 적은 오차율을 보이는 것으로 나타났다. 무작위로 추출된 20개의 교육시설 사례에 대한 예측 결과, 평균 오차율은 7.48%를 보였다. 하지만 이는 다중회귀분석 모델의 평균 오차율(5.68%)보다는 상대적으로 높은 수치였다. 두 모델간의 유의성 판단을 위해 분산분석을 수행한 결과, 두 모델의 평균 차이가 통계적으로 유의하지 않은 것으로 나타났다. 또한 SVM모델이 다중회귀모델에 비해 오차율 및 오차범위가 크게 발생하므로 교육시설 공사비 예측에는 SVM모델보다 다중회귀모델이 더 적합할 것으로 사료된다. 따라서 SVM모델의 예측 정확도 향상 및 적용가능성을 높이기 위해서는 교육시설 프로젝트에 대한 충분한 사례의 확보가 필요할 것이다. 이를 바탕으로 추가적인 연구를 통해 SVM을 이용한 교육시설 초기공사비 예측 모델의 정확성을 높이는 방안을 모색해야 할 것이다.

또한 본 연구에서는 경기도교육청에서 2004년에서 2007년 사이에 발주한 교육시설 프로젝트의 실적데이터만을 대상으로 예측 모델을 구축하였다. 하지만 건축 프로젝트 초기단계에서 신속하고 정확한 공사비 예측 및  위해서는 다양한 시설 및 지역의 데이터를 축적하여 연구를 진행하는 것이 필요할 것으로 사료된다.

## 참고문헌

1. 김광희 외, 사례기반추론 기법을 이용한 공동주택 초기 공사비 예측에 관한 연구, 대한건축학회 논문집, 제20권, 제5호, pp.83-92, 2004
2. 김광희 외, 유전자 알고리즘에 의한 신경망 구조의 최적화를 이용한 공동주택의 초기 공사비 예측에 관한 연구, 대한건축학회논문집(구조계), 제 20권, 제2호, pp.81-88 , 2004
3. 김광희 외, 공동주택 공사비 예측 정확도 비교에 관한 연구, 대한건축학회논문집(구조계), 제20권, 제5호, pp.93-102, 2004
4. 김광희 외, 신경망과 유전자알고리즘을 이용한 공사비예측 모델의 예측정확도 비교에 관한 연구, 대한건축학회논문집(구조계), 제22권, 제3호, pp.111-118, 2006
5. 김유일 외, 신경망과 SVM을 이용한 주가지수예측의 비교, 인터넷전자상거래연구, 제4권, 제3호, pp.221-243, 2004
6. 김진원 외, 회귀분석을 이용한 교육시설의 공간계획에 따른 공사비 예측 모델에 관한 연구, 대한건축학회논문집, 제27권, 제10호, pp.103-110, 2011
7. 박우열 외, 서포트 벡터 회귀분석을 이용한 공동주택 공사비 예측에 관한 연구, 대한건축학회 논문집(구조계), 제23권, 제4호, pp.165-172, 2007
8. 박우열 외, Support vector machine을 이용한 개산견적 평가모델에 관한 연구, 대한건축학회논문집(구조계), 제23권, 제4호, pp.191-198, 2005
9. 박우열 외, Support vector machine을 이용한 흙막이 공법 선정모델에 관한 연구, 건설관리, 제7권, 제2호, pp.118-126, 2006
10. 박종석, 데이터마이닝에서 서포트 벡터 회귀분석과 신경망 분석 기법의 비교연구, 동국대 석사학위논문, 2006
11. 손재호 외, 신경망을 이용한 교육시설 BTL 사업의 공사비 분석 및 예측에 관한 연구, 대한건축학회논문집 (구조계), 제24권, 제6호, pp.135-142, 2008
12. 안성훈 외, 전문가지식을 활용한 공동주택 초기단계 공사비 예측에 관한 연구, 대한건축학회 논문집(구조계), 제21권, 제6호, pp.81-88, 2005
13. 이진선 외, 대용량 분류에서 SVM과 신경망의 성능 비교, 정보처리학회논문집 B, 제12B권, 제1호, 2005
14. 정영미 외, SVM 분류기를 이용한 문서 범주화 연구, 정보관리학회지, 제17권, 제4호, pp.229-248, 2000
15. 한국건설기술연구원, 2012년 9월 건설공사비지수 동향, 2012
16. Burges CJC, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery, Vol.2, No.2, pp.121-167, 1998
17. Debasish, B et al. Support Vector Regression, Neural Information Processing-Letters and Review, Vol.11, No.10, pp.203-224, 2007
18. Dumais, S et al. Inductive learning algorithms and representations for text categorization, Proceedings of ACM-CIKM 98, pp.148-155, 1998
19. Kim GH et al., Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning, Building and Environment, Vol.39, pp.1235-1242, 2004
20. Kim S, Support vector machine model to select Exterior Materials, Journal of the Korea Institute of Building Construction, Vol.11, No.3, pp.238-246, 2011
21. Smola AJ et al, A tutorial on support vector regression, Neuro COLT Technical Report TR-98-030, Royal Holloway College University of London, UK, 1998
22. Vapnik V. The Nature of Statistical Learning Theory, 2nd Ed, Springer, New York, p.1-314, 1995

(논문투고일 : 2012.10.26, 심사완료일 : 2012.12.04, 게재확정일 : 2012.12.14)