적은 수의 표본에 기초한 흙막이 공법선정 방법에 대한 기초연구

A Framework for Developing a Method for Selecting a Retaining Wall System Using a Small Number of Samples

최 명 석* 이 강**
Choi, Myung Seok Lee, Ghang

요 약

지난 10여 년간 건설 분야에서 데이터 마이닝(data mining) 기법을 이용하여 실적 데이터의 패턴을 찾고 이를 바탕으로 다양한 예측 모델을 개발하려는 연구들이 다수 진행되어 왔다. 그러나 데이터 마이닝 기법의 경우, 일반적으로 수천 또는 수만 개의 데이터를 사용할 것을 추천하고 있으나, 건설 분야의 선행연구들을 살펴보면 데이터 마이닝에 사용된 데이터의 수가 최대 이백 개 표본 정도에 그치고 있다. 그 결과 데이터 마이닝 기법을 이용하여 선행연구의 방법 및 결과를 재현하려고 하여도 같은 결과를 얻기 힘들고, 아예 패턴추출이 힘들거나 도출된 패턴에 오류가 있는 경우도 있다.

본 연구는 소수의 표본수에 기초하여 보다 신뢰성 있는 공법선정 모델의 개발에 관한 대안적인 방법을 모색하고자 한다. 먼저 건설 분야에서 데이터 마이닝 기법을 활용한 연구의 범위에 대해 살펴보고, 특히 흙막이 공법선정 모델 구축에 관한 선행연구 고찰을 통해 소수의 데이터에 의존함에 따라 발생하는 문제점들을 알아보고, 이를 극복할 수 있는 방안에 대해 논의 하고자 한다.

키워드: 데이터 마이닝, 흙막이 공법 선정, 예측 모델

1. 서 론

1950년대 이후 컴퓨터의 급속한 발달로 통계학을 기초로 데이터 마이닝(data mining)을 통해 대용량의 데이터에서 숨겨진 유용한 정보와 관계를 탐색하고 모형화하는 통계적 기법[1]들이 다양한 분야에서 활용되고 있다. 최근 건설 분야에서도 실적 데이터의 다양한 통계적 분석을 통해 구축된 예측 모델을 활용하여 프로젝트 초기 단계에서 제한된 정보에 기초한 적절한 대응방안 도출 및 의사결정에 도움이 되는 정보를 획득하고자 하는 연구들이 다수 수행되고 있다 (Yau와 Yang 1998, Yau 외 1999, Koksal과 Arditi 2004, 안성훈과 강경인 2005, 김재엽 외 2006, 김태훈 외 2007).

하지만 대부분의 선행 연구에서 활용된 데이터[2]의 수는 최대 수백을 넘지 않는다. 일반적으로 데이터 마이닝 기법을 통해 의미 있는 결과를 도출하기 위해 요구되는 데이터의 수는 수천 혹은 수 만개 정도이다 (Berry와 Linoff 2000). 따라서 건설 분야에서 데이터 마이닝 기법을 활용한 예측 모델 구축에 사용되는 수백 개의 자료는 극히 적음을 알 수 있다. 물론 대형 마트에서 바코드를 통해 지속적으로 기록되는 수만 건의 소비자 상품거래(transaction) 데이터를 바탕으로 연관규칙(association rule)을 발견하는 경우와 건설 산업에서 공사비 예측을 위해 어렵게 소수의 실적 공사비 데이터를 수집하는 것은 경우가 다르다. 그러나 건설 산업에서 수입될 수 있는 제한된 데이터의 수에 의존하여 연구를 진행함에 따라, 데이터 마이닝 기법을 이용한 일부 선행연구에서 자동으로 추출된 패턴에 오류가 발견되거나 그 결과를 재현하기 어려운 경우가 나타난다.

따라서 본 연구는 자동화된 데이터 마이닝 기법에 전적으로 의존하는 대신 비교적 적은 수의 표본에서도 신뢰성이 높은 분석결과를 얻을 수 있는 통계적 분석기법의 활용방안을 모색하고자 한다. 특히 흙막이 공법선정 모델에 관한 선행 연구를 중심으로 소수의 데이터를 활용함에 따라 야기되는 문제점들을 알아보고, 이를 극복하기 위한 방안에 대해 논의 하고자 한다.

2. 국내외 선행연구 고찰

2.1 데이터 마이닝 기법을 활용한 연구 범위

* 일반회원, 연세대학교 건축공학과 석사과정, isedyou@hanmail.net
** 종신회원, 연세대학교 건축공학과 조교수, Ph.D., 교신저자, glee@yonsei.ac.kr

본 연구는 국토해양부/한국건설교통기술평가원의 건설핵심기술연구사업 「공기단축형 복합구조시스템 건설기술」 (05R&D/건설핵심D02-01)의 연구비 지원에 의한 연구의 일부임.

1) 통계적 기법은 통계학(statistics), 기계 학습(machine learning), 데이터 마이닝(data mining)과 관련하여 자료를 서술하고 나아가 추정 및 예측에 활용되는 기법을 포함하며, 본 연구에서는 데이터 마이닝으로 총칭한다.

2) 본 연구에의 데이터는 기계 학습에 사용되어 모델을 구축하는데 활용된 데이터뿐만 아니라 모델의 평가에 활용되는 모든 데이터를 포함하는 것을 의미한다.

건설 분야에서 데이터 마이닝 기법을 이용한 예측 모델 구축에 관련된 선행연구는 흙막이 벽체 선정, 건설 중재 결과 성취 여부, 공사비 예측 등 매우 다양한 범위에서 적용되어 왔다. 본 연구에서는 표 1에 제시된 연구들을 중심으로 데이터부족으로 인한 문제점을 살펴보도록 한다.

표 1. 국내외 통계적 분석을 활용한 주요 선행연구 예시

저자	목표변수	주요 설명변수	분석 기법	사례수
Yau 외(1999)	12수준의 흙막이 벽체 종류 선정	시공 현장 조건에 해당하는 10개 변수	규칙 추론	254
Yau와 Yang (1998)	12수준의 흙막이 벽체 종류 선정	시공 현장 조건에 해당하는 8개 변수	사례기반 추론	254
김재엽 외(2003)	4가지의 흙막이 지보공 공법	8가지의 현장 여건	인공 신경망	223
Koksal과 Arditi (2004)	3수준의 건설회사 현황 예측	조직체계, 인적자원, 전략적 방침에 해당하는 21개 변수	다항 로지스틱 회귀분석	52
Yang (2004)	11수준의 흙막이 벽체 종류 선정	시공 현장 조건에 해당하는 9개 변수	분석기법 결과종합	254
안성훈과 강경인 (2005)	지하주차장 공사비	연면적, 외벽길이, 기초형태 등의 8개 변수	선형 회귀분석	26
Yiu 외(2006)	4가지의 건설 중재 결과의 성취 여부	9가지의 건설 중재 전략	이항 로지스틱 회귀분석	32
김태훈 외(2007)	각 3가지의 바닥 거푸집 및 외부벽체 거푸집	건물층수, 층당공기, 구조형식	의사결정 트리	61

Koksal과 Arditi(2004)는 비재정적(nonfinancial) 변수의 고려를 통해 현재 건설 회사가 건전한지 혹은 쇠퇴기에 있는지를 예측하기 위한 모델을 구축하였다. 이를 위해 법정관리 경험이 있는 건설 회사를 대상으로 자료를 수집하였는데, 법정관리를 신청했던 경험이 있는 건설 회사로부터 데이터를 수집하기란 매우 어려운 일이었다. 결국 11개 회사만이 응답하였으며(응답률 8%) 정상적인 회사로부터 수집된 41개를 포함해 총 52개의 응답 자료만이 수집되었다. 이 중 결측치를 포함한 자료를 제외한 46개만이 분석에 사용되었다. 이렇게 수집된 자료 또한 도수분포 측면에서 쇠퇴 초기 단계에 심하게 편중된 자료였다. 이러한 클래스 불균형(class imbalance)은 정확한 분류 경계를 설정하는데 장애가 되어, 종종 대부분의 데이터를 다수 범주로 분류하는 모델을 구성하게 된다(Hosmer와 Lemeshow 2000). 최종적으로 수립된 모델은 전반적으로 80.4%의 예측 정확도를 나타냈다. 그러나 자세히 살펴보면 쇠퇴 초기 단계(다수 범주)에 대해서는 100%의 분류 정확도를 보이지만, 나머지 두 소수 범주에 대해서는 분류 정확도가 각 28.6%와 33.3%에 그치는 문제점을 표출하였다. Kubat 외(1997)는 다수 범주와 소수 범주의 정확도를 모두 고려하기 위해서 기하평균(geometric mean)을 이용하였으며, 이를 적용할 경우 예측도가 45.7%가 된다. 또한 클래스 불균형 문제를 다루기 위한 다양한 샘플링 및 앙상블 기법들에 대한 고려가 미흡했다. 제시된 모델의 검증에 있어서도 이상적으로는 모델 수립에 사용되지 않았던 독립적인 검증용 데이터를 사용해야 하지만 응답 자료의 부족으로 이미 모델 수립에 사용된 자료에서 3개의 사례만을 무작위로 선택하여 검증하였다. 수립된 모델의 효용성을 올바로 검증하지 못하는 문제는 Yau와 Yang(1998), Yau 외(1999)의 흙막이 벽체 선정 모델 구축에 관한 연구에서도 나타난다. 254개의 과거 터파기 공사 시공사례를 바탕으로 사례기반추론(case-based reasoning)과 규칙추론(rule induction) 기법을 이용하여 현장 여건에 적합한 흙막이 벽체 선정 모델을 구축하였다. 하지만 제시된 모델의 효용성은 4개의 검증사례를 통한 예측정확도에만 의존하여 판단되었다. Yau 외(1999)의 연구에서 총 4개의 검증용 사례들을 바탕으로, 실제로 적용된 공법과 예측모델에서 제시된 공법의 일치 여부를 비교하였다. 하지만 4개의 검증용 사례에만 의존한 결과인 데다가 복수의 예측공법 중 하나가 맞으면 예측이 맞은 것으로 결론을 지어 연구결과의 신뢰도가 매우 낮다.

인공신경망의 경우 Berry와 Linoff(2000)는 학습 데이터에 대한 과적합(overfitting)없이 효과적으로 학습시키기 위한 최소한의 학습 데이터 수를 다음과 같이 제안하고 있다. 하나의 은닉층과 하나의 출력 노드를 가지는 상황에서 h개의 은닉노드와 n개의 입력노드에 따라 h*(n+1)+1의 수만큼의 가중치(연결의 수)가 필요하며, 각 가중치마다 최소 10개의 학습 데이터가 필요하다. 또한 범주형 목표변수를 추정할 경우에는 각 범주마다 상기 제시된 수만큼의 학습 데이터가 필요하다. 흙막이 지보공 공법 선정을 위한 선행 연구에서 8개의 입력노드와 10개의 은닉노드, 그리고 4개의 목표 범주를 가진 인공신경망을 구축하였으며, 위의 제안을 적용한다면 필요로 되는 학습 데이터의 수는 3640개 정도이나 실제 연구에서는 96개의 학습 데이터가 사용되었다. 물론 인공신경망은 사용자가 최종 결과물을 도출하는 내부 과정을 알 수 없는 일명 블랙박스(black-box characteristic) 형태로 되어 있어서 수립된 모델의 효용성을 판단하기 위해서는 학습데이터 수는 관계가 없으며 결과적으로 나타난 예측 정확도에 의존할 수밖에 없다. 최종적으로 수립된 학습오차가 가장 작은 인공신경망 모델에서는 96개의 학습 데이터의 경우 99%의 예측 정확성을 보였으며, 47개의 검증용 데이터에서는 77%의 정확성을 보였고, 총 91.6%의 정확성을 보인 것으로 보고되어, 샘플수가 적은 것이 연구결과에 영향을 미쳤다고 보기는 어렵다. 그러나 저자들이 비슷한 수의 다른 표본을 가지고 인공신경망 기법을 적용한 결과 의미있는 결과를 재현하기가 어려웠다.

데이터의 부족은 다양한 시나리오에서 모델을 수정해 나가며 최적의 예측 모델을 선정하는데 어려움을 주기도 한다. 공동주택의 지하주차장 공사비 예측을 위한 선행 연구에서는 26개의 지하주차장에 대한 실적자료를 바탕으로 회귀분석을 이용한 모델과 부위별

단가를 이용한 모델을 각각 구성하여 공사비 예측 성능을 비교하는 연구를 수행하였다. 이 중 회귀분석을 이용한 공사비 예측모델을 구성함에 있어서 연속형 변수인 연면적, 지하층수, 외벽길이, 램프개수, 일체식 연면적과 2수준의 명목형 변수인 락앙카적용, 주차장 형태, 기초형태의 총 8개의 설명변수들을 이용하였다. 이 중에서 단계별 선택법(stepwise)을 통해 결정계수(R^2)가 가장 큰 단계에서 선택된 연면적, 일체식 연면적, 램프개수, 기초형태의 4가지 설명변수로 구성되는 최종 회귀 모델을 선정하였다. 본 논문의 저자는 기존 논문에 제시된 26개의 지하주차장 데이터를 바탕으로 각 설명변수들 간의 상관관계 및 실질적 의미와 다중공선성 여부를 살펴보고 최종적으로 연면적과 일체식 연면적의 두 가지 설명변수로 구성된 회귀모델을 도출하였다. 기존 연구에서 모델의 평가를 위한 기준으로 사용하고 있는 실제공사비와 예측값과의 오차율에 따르면, 기존의 4개의 설명변수로 구성된 회귀모델의 경우 13.02%의 오차율을 보였다. 본 연구에서 2개의 설명변수로만 구성된 회귀모델은 이보다 작은 9.57%의 오차율을 보였다. 모델의 간명성(parsimony)이나 실제값과의 오차율 측면에서 새로 수립된 회귀모델이 더 나은 모델이라고 판단할 수도 있으나, 분석 및 평가에 사용된 사례가 극히 적은 상황에서 이는 우연에 따른 판단(capitalization on chance)이 될 수도 있다.

규칙 추론(Rule induction)이나 의사결정 트리(decision tree)와 같은 기법들은 데이터 마이닝의 추론과정이 if-then형식의 규칙이나 도식화된 패턴으로 표현되므로 귀납된 결과를 전문가가 본인의 지식을 바탕으로 검토하는 것이 가능하다. Yang(2004)의 연구에서 목표변수인 11개의 흙막이 벽체 종류 중 규칙 추론을 통해 제시된 auger boring pile에 관한 의사결정 규칙을 통해 귀납된 규칙의 실질적 의미에 대해 살펴보면 몇 가지 모순점을 발견할 수 있다(그림 1 참고).

그림 1 . 규칙추론을 통해 귀납된 규칙 예시(Yang 2004, p42)

(1) If Groundwater<0.550 And Soil_Type is Sandy_Gravel And Location is TaipeiCity Then RWS_Auger_Boring_Pile is Yes	(2) If Groundwater<0.550 And Soil_Type is Sandy_Gravel And Location is Taipei Then RWS_Auger_Boring_Pile is No
(3) If Groundwater>=0.550 And Soil_Strength_Firm is Yes And Soil_Strength_V_Soft is No And Groundwater<1.625 And Location is TaipeiCity Then RWS_Auger_Boring_Pile is No	(4) If Groundwater>=0.550 And Soil_Strength_Firm is Yes And Soil_Strength_V_Soft is No And Groundwater<1.625 And Location is Taipei Then RWS_Auger_Boring_Pile is Yes
(5) If Groundwater>=0.550 And Soil_Strength_Firm is Yes And Soil_Strength_V_Soft is Yes Then RWS_Auger_Boring_Pile is Yes	

규칙 (1)과 (2)는 지역(location)이 "Taipei City"와 "Taipei"로 서로 다른 것을 제외하고 나머지 조건들이 모두 같지만, 규칙 (1)은 "auger boring pile"의 적용을 추천하고 있으며 규칙 (2)는 그렇지 않다. 규칙 (3)과 (4)도 지역이 다른 것을 제외한 나머지 조건은 동일하지만 최종결과는 규칙 (1)과 (2)를 통해 제시된 것과 서로 상반된다. 규칙 (5)에서는 토질 강성(soil strength)이 상호 배타적인 조건인 "단단함(firm)"과 "연약함(soft)"이 모두 Yes일 경우 auger boring pile을 추천하고 있다. 또한 지하수위 분류의 기준이 된 0.55m와 1.625m는 실무적으로 봤을 때 공법선정에 영향을 주기에는 너무 작은 수치이다.

2.2 선행연구에 나타난 제약사항 요약

데이터의 수가 적은 경우 다음과 같은 문제가 발생하는 것으로 나타났다. 첫째로 통계적 분석에서 요구하는 가정(assumption)을 충족시키지 못함(예: 표본이 정규성이나 등분산성을 이루지 못함)에 따라 정확한 통계적 유의도를 판단하지 못하거나, 일부의 전형적인 사례에만 잘 작동하는 불안정한 모델이 구축될 수 있다. 특히 의사결정나무와 같이 인간이 인지 가능한 형태로 결과가 도출되는 경우 결과의 비논리적인 부분도 발견할 수 있었다. 둘째로 모델을 구축하는 다양한 시나리오(예: 인공신경망에서 은닉층의 노드 수 결정, 사례기반 추론에서 각 인덱스의 가중치 부여, 회귀분석에서 설명변수의 선택 등) 속에서 주어지는 상이한 결과들을 수정하며 최선의 모델을 선정하는 것이 어렵다. 추정 되는 모델 평가의 기준값이나 지극히 소수의 검증사례를 가지고 평가된 예측 정확도를 신뢰할 수 없는 경우이다. 물론 모집단 전체를 대상으로 추정하지 않는 한, 어떠한 모델도 궁극적으로 최선의 모델이라고 말 할 수 없겠지만, 제한된 데이터에 기초한 모델 평가의 기준값이나 예측 정확도에 의존한다면 수립된 모델의 우수한 예측 성능이나 최선의 모델이라는 판단이 우연의 일치에 따른 것일 수 있다. 셋째로 제한된 데이터에 의존함으로써 발생될 수 있는 문제들에 대한 해결 및 모델의 예측 성능의 향상에 대안적인 방법들과 모델의 평가 기준에 대한 면밀한 검토 없이 자동화된 기계 학습에 의존하는 것이다. 이 경우 분석 과정에서 모델 수립자 혹은 목표로 되는 분야에 해당 분야 전문가의 실질적인 지식이 반영되기 어렵다.

3. 결론 및 제안 사항

최근 건설 분야에서 과거 실적 데이터의 통계적 분석을 통해 의사결정에 도움이 되는 유용한 정보를 도출하기 위한 시도가 다양한 방면에서 연구되고 있다. 하지만 대부분의 선행연구에서 공통적으로 직면한 문제는 건설 산업의 특성 혹은 예측 변수의 특성상 통계적 분석을 통해 신뢰할만한 정보를 도출할 수 있는 충분한 양의 실적 데이터 수집 및 통계적 분석에서의 활용이 매우 어렵다는 것이다. 선행연구에서 소수의 과거 실적 데이터에 의존함에 따라 야기될 수 있는 문제들의 해결에 대한 다양한 시도나 통계적 분

석 과정 및 모델의 평가에 대한 면밀한 검토가 부족할 경우 수립된 모델의 불안정성 및 귀납된 결론의 오류가 발생하며, 최선의 모델을 선정하고 수립된 모델을 평가하는 것이 우연에 따른 결과일수도 있음을 살펴보았다.

이러한 문제를 조금이나마 해결하기 위해 본 연구에서는 다음의 4가지 사항을 제안하고자 한다. 첫째로 통계모델에서 목표변수와 설명변수 사이의 관계에 대한 실질적 의미를 살펴보아야 할 것이다. 통계적으로 유의하다고 반드시 실제 의미가 있다고 볼 수 없으며, 반대로 실질적인 의미는 중요하나 통계적으로는 유의하지 않다고 판단될 수 도 있다. 분석 이전에 선행연구 고찰 및 전문가 의견을 수렴해 실질적인 설명변수를 정의하나, 통계적 분석의 결과에 대한 실질적 의미에 대한 검토는 대부분 이루어지지 않고 있다. 또한 통계적 분석 결과와 실질적 의미에 차이가 있을 경우 수립된 모델을 수정하거나 기존 설명변수에 대한 재해석이 필요할 수 있다. 둘째로 실질적 의미와 통계적 유의성이 차이가 있을 수 있다는 전제하에 수립된 모델의 실질적인 의미를 검증 할 수 있도록 분석결과가 명시적으로 이해될 수 있는 형태로 표현하는 것이 유리하다. 예를 들어 데이터의 수가 제한적인 상황에서 목표변수와 설명변수와의 관계가 알려져 있지 않아 패턴의 특성을 발견하여 일반화에 유리한 인공신경망을 사용한다면, 데이터에 과적합 되거나 혹은 우연에 따른 결과가 도출될 수도 있다. 따라서 목표변수와 설명변수의 관계를 모른다면 통계적 분석을 통해 귀납되는 관계를 다시 실질적으로 검토해보는 일련의 과정을 통해 근본적인 관계를 정의하는 것이 중요하다. 특히 일반화 가능한 논리적인 결과를 도출할 수 있는 양질의 다수 데이터가 확보되지 않았다면 더욱 그러할 것이다. 셋째로 주어진 데이터를 바탕으로 최선의 결과를 얻을 수 있도록 대안적인 방법들과 통계적 분석과정에 대한 면밀한 검토가 필요하다. 예를 들어 수집된 데이터의 클래스가 불균형이라면 샘플링(sampling)이나 앙상블(ensemble)기법의 활용을 통해 수립되는 모델의 예측 성능을 향상시키기 위한 시도가 필요할 것이다. 넷째로 지속적으로 분석에 활용될 수 있는 데이터를 확보하기 위한 시도가 필요하다. 수립된 예측 모델을 웹기반의 시스템으로 구축하여 사용자로부터 해당 분야의 데이터 및 전문가 의견을 수집하고, 이를 반영한다면 보다 좋은 성능의 예측 모델(시스템) 구축이 가능할 것이다.

참고문헌

1. 강필성, 조성준, (2006) "데이터 불균형 해결을 위한 Under-Sampling 기반 앙상블 SVMs." 한국경영과학회 춘계공동학술대회.
2. 김재엽, 서장우, 강경인. (2003), "신경망을 이용한 흙막이 지보공공법 선정모델 개발에 관한 연구." 대한건축학회 논문집(구조계), 19(5), pp. 121-128.
3. 김태훈, 신윤석, 이웅균, 강경인, (2007). "의사결정나무를 이용한 초고층 건축공사 거푸집 선정 지원 모델." 대한건축학회 논문집(구조계), 23(11), pp. 117-124.
4. 안성훈, 강경인, (2005). "공동주책의 지하주차장 공사비 예측 모델에 관한 연구." 대한건축학회 논문집(구조계), 21(5), pp. 135-142.
5. Berry, M. J. A., and Linoff, G. (2000). Mastering data mining : the art and science of customer relationship management, Wiley Computer Pub., New York.
6. Hosmer, D. W., and Lemeshow, S. (2000). Applied logistic regression, Wiley, New York.
7. Koksal, A., and Arditi, D. (2004). "Predicting Construction Company Decline." Journal of Construction Engineering and Management, 130(6), 799-807.
8. Kubat, M., Holte, R., and Matwin, S. "Learning when Negative Examples Abound." Proceedings of the 9th European Conference on Machine Learning.
9. Sheu, H. B. (1996). "Application of Expert Systems and Neural Networks for Retaining Wall System Selection," National Central University, Taiwan.
10. Yang, J. B. (2004). "Hybrid AI system for retaining wall selection." Construction Innovation, 4(1), 33-52.
11. Yau, N. J., Yang, J. B., and Hsieh, T. Y. (1999). "Inducing Rules for Selecting Retaining Wall." Construction Management and Economics, 17, 91-98.
12. Yau, N.-J., and Yang, J.-B. (1998). "Applying case-based reasoning technique to retaining wall selection." Automation in Construction, 7(4), 271-283.
13. Yiu, T. W., Cheung, S. O., and Mok, F. M. (2006). "Logistic Likelihood Analysis of Mediation Outcomes." Journal of Construction Engineering and Management, 132(10), 1026-1036.

Abstract

In the past decade, various data mining techniques have been used in construction engineering as a means to make informed decisions through the aid of useful knowledge discovered from historical data. Researchers in the construction domain are often confronted with a challenge to derive a meaningful conclusion with a limited sample of data. However, when the data size is small, the proposed results are often illogical. Even if the derived results are technically flawless, sometimes it is difficult to reproduce these results by using the same analysis method when a different set of data is used. This paper reviews some problems that stem from limited data size, and discusses several recommendations for dealing with these problems.

Keywords : Data mining, Retaining wall selection, Prediction