교원교육, 한국교원대학교 교육연구원
Korean Journal of Teacher Education
2017, Vol. 33, No. 1, pp. 43-56
https://doi.org/10.14333/KJTE.2017.33.1.43

기계학습을 통한 TIMSS 2011 중학생의 수학 성취도 관련 변수 탐색

유 진 은*

2016. 11. 16.(접수)/2017. 01. 04.(1심통과)/2017. 01. 06.(최종통과)

<< 요 약 >>

본 연구는 기계학습적 접근법인 LASSO 기법을 우리나라 TIMSS 2011 중학교 2학년 자료에 적용하였다. TIMSS의 100개의 설명변수를 모형에 모두 투입하여 22개 변수를 선택하였을 때, 이 모형의 예측정확도는 약 80%였다. 학생의 수학적 자기효능감, 수학에 대한 태도, 어머니의 교육 수준, 그리고 가정 보유 장서 수와 같은 가정의 교육자원 변수가 학생의 수학 성취 수준에 영향을 미치는 것으로 나타났으며, 이는 기존 연구 결과와 일치하였다. 본 연구에서 학생의 수학 성취수준과 관련과 있다고 새롭게 탐색된 변수로 수학숙제 시간, 학생의 과학적 자기효능감, 과학숙제 부여 빈도 등이 있었다. 연구 함의 및 향후 연구 주제 또한 논의되었다.

주제어: 기계학습, LASSO, 축소추정법, TIMSS, 수학성취도

Ⅰ. 서 론

지난 3월 바둑기사 이세돌이 알파고(AlphaGo)에게 4:1로 패한 후, 사람들은 커다란 충격을 받았다. 이미 인공지능이 인간을 제치고 우승한 이력이 있기는 하였다. 1997년에 IBM이 개발한 인공지능 'Deep Blue'가 인간 체스 챔피언인 Garry Kasparov를 이겼고, 2011년에 역시 IBM의 'Watson'이 퀴즈쇼 Jeopardy에서 인간 대표를 제치고 백만 불의 상금을 획득했던 것은 인공지능 역사에서 중요한 이정표로 회자되고 있다. 그러나 경우의 수가 무한에 가까운 바둑의 경우 수십

* 한국교원대학교 : jeyoo@knue.ac.kr

년은 지나야 인공지능이 인간 고수와 대적할 수 있을 것이라고 전문가들은 예견했다. 그런데 그로부터 불과 5년 후 인공지능이 18번이나 세계 바둑챔피언을 지냈던 인간 최고수를 이김으로써, 학계뿐만 아니라 일반인들의 인공지능에 대한 관심이 급격하게 높아졌다.

알파고와 같은 인공지능을 가능하게 하는 기술 중 핵심이 되는 부분이 기계학습(machine learning)이다. 간단하게 말하자면, 기계학습은 알고리즘을 통해 기계(컴퓨터)를 학습시킴으로써 최적의 결과를 찾아내도록 하는 방법으로, 컴퓨터공학의 패턴 인식 프로그램이 그 대표적인 예가 될 수 있다. 더 자세하게 설명하자면, 알파고는 기계학습 기법 중 'Monte Carlo tree search' 알고리즘과 신경망(Neural Networks)의 한 형태인 'Policy Networks'와 'Value Networks'를 이용하였다(Silver et al., 2016).

신경망 모형 구축 시 고려할 사항 중 하나는 과적합(overfitting) 문제다. 신경망 모형에서 추정해야 하는 가중치들이 기하급수적으로 증가할 수 있기 때문이다. 이러한 과적합 문제를 처리하기 위한 방법 중 주로 이용되는 방법이 정규화(regularization) 기법이다. 정규화 기법 중 LASSO(Least Absolute Shrinkage and Selection Operator)와 같은 축소추정법이 최근 통계학을 비롯한 여러 학문 분야의 연구 주제로 주목을 받고 있다(Yoo, 2016).

반면, 초중등 학생의 학업성취도 연구에 이러한 기법이 적용된 사례는 찾기 힘들다. 대학생까지 대상으로 하며 데이터 마이닝 기법 전반으로 범위를 넓혀서 검색할 때, 대학교 컴퓨터 교양교육(Kim, 2012; Kim, 2013)과 특성화 고등학교(Kim & Yong, 2014)의 교육성과를 알아보기 위하여 의사결정나무 모형을 이용한 논문 세 편을 찾을 수 있을 뿐이었다.

본 연구는 학업성취도 연구에 기계학습 기법을 적용한 처음 연구 중 하나라고 할 수 있다. 더 자세히 말하자면, LASSO를 TIMSS(Trends in International Mathematics and Science Study) 2011 한국 자료에 적용함으로써 우리나라 중학생들의 수학 성취도에 영향을 미치는 변수들을 탐색하였다. TIMSS는 수백 개에 이르는 변수를 제공하기 때문에 변수 선택에 강점을 보이는 LASSO 방법을 적용하기 적합한 자료다.

Ⅱ. TIMSS 선행연구

TIMSS는 세계 여러 나라 4학년과 8학년 학생들의 수학 및 과학성취도와 그 맥락변인들을 1995년부터 4년마다 측정해 왔다. TIMSS 2007과 TIMSS 2011 자료를 분석한 최근 연구에서 학생들의 수학/과학에 대한 태도(Azina & Halimath 2012; Ng, Lay, Areepattamannil, Treagust, & Chandrasegaran, 2012; Tsai & Yang, 2015; Winnaar, Frempong, & Blignaut, 2015)와 수학/과학에 대한 자기효능감(Azina & Halimah, 2012; Engel, Rutkowski, & Rutkowski, 2009; Sulku &

Abdioglu, 2015; Tsai & Yang, 2015; Winnaar et al., 2015)이 수학/과학 성취도 예측 모형에 자주 이용된 변수였다.

SES(Social Economic Status; 학생의 사회경제적 지위)가 학업성취도와 상관이 높다고 알려져 있다. 여러 TIMSS 연구에서 가정에서 보유하고 있는 책의 양(Azina & Halimah, 2012; Engel et al., 2009; Kareshki & Hajinezhad, 2014; Matsuoka, 2014; Sulku & Abdioglu, 2015; Tsai & Yang, 2015), 가정에서의 컴퓨터 보유 유무(Azina & Halimah, 2012; Kareshki & Hajinezhad, 2014; Sulku & Abdioglu, 2015), 가정에서의 다른 기기(예: 책상, 인터넷 연결 등) 보유 유무(Azina & Halimah, 2012; Matsuoka, 2014), 어머니의 교육 수준(Cheng, 2014; Mills & Holloway, 2013) 등으로 SES를 간접적으로 측정하고 학업성취도 예측 모형에 이용하였다.

학생의 나이(Kareshki & Hajinezhad, 2014; Winnaar et al., 2015), 성별(Azina & Halimah, 2012; Engel et al., 2009; Ng et al., 2012; Kareshki & Hajinezhad, 2014; Sulku & Abdioglu, 2015; Tsai & Yang, 2015)과 같은 인구통계학적 변수뿐만 아니라 가정에서 쓰는 언어(Azina & Halimah, 2012; Kareshki & Hajinezhad, 2014; Ng et al., 2012; Tsai & Yang, 2015) 또한 연구되었다. 학교 일에 대한 부모의 관여(Kareshki & Hajinezhad, 2014; Sulku & Abdioglu, 2015; Winnaar et al., 2015), 학교폭력 및 학교왕따(Azina & Halimah, 2012; Engel et al., 2009; Winnaar et al., 2015), 학교에서의 컴퓨터 이용 등도 학업성취도를 예측하는 설명변수였다(Azina & Halimah, 2012; Sulku & Abdioglu, 2015).

Ⅲ. LASSO 기계학습법

변수가 많은 자료로 모형 구축 시 변수 선택은 매우 중요한 이슈다. 축소추정법(shrinkage estimation methods)은 계수에 벌점(penalization)을 부과함으로써 과적합 문제를 해결하고자 한다. 본 연구에서는 축소추정법 중 하나인 LASSO(Least Absolute Shrinkage and Selection Operator)를 이용하였다. LASSO는 1996년 Tibshirani가 발표한 방법이다. 능형회귀(ridge regression) 또한 축소추정법이지만 변수 선택은 하지 않는 데 비해, LASSO는 계수 추정 및 변수 선택을 동시에 수행한다는 장점이 있다(Hastie, Tibshirani, & Friedman, 2009). LASSO 추정식은 식 (1)과 같다:

$$\hat{\beta} = \arg min\left\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P}x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{P}|\beta_j|\right\} \quad (1)$$

식 (1)에서 λ가 벌점모수(penalty parameter)가 되며, λ값에 따라 축소추정을 어느 정도로 할

것인지 결정된다. 벌점모수 λ는 0과 무한대 사이의 값을 가질 수 있는데, λ값이 클수록 회귀계수가 0에 수렴을 하고, λ가 작아질수록 회귀계수가 최소제곱법의 추정치와 가까워진다. 즉, 벌점모수인 λ를 크게 하여 선택된 회귀계수 추정값을 0으로 만드는 경우 자동적으로 변수 선택이 이루어진다. 따라서 λ의 크기를 결정하는 것이 LASSO를 통한 계수 추정에 중요한 부분이 되며, 이때 일반적으로 교차타당화(cross-validation: CV)가 이용된다(Hastie et al., 2009; Yoo, 2016).

Ⅳ. 연구방법

1. 분석 자료

본 연구에서는 TIMSS(Trends in Mathematics and Science Study) 2011의 우리나라 중학교 2학년 학생 자료를 분석하였다. 학생의 수학 성취수준을 반응변수로 하고 나머지 변수들을 설명변수로 하여 어떤 변수들이 수학 성취도와 연관되어 있는지를 탐색하는 것이 분석의 목적이었다. 총 5166명의 학생에 대하여 397개 변수가 있었으나, 그 중 많은 변수들이 전체 무응답이었다. 이를테면 과학의 하위 교과인 물리, 화학, 생물, 지구과학 관련 문항들의 경우 우리나라 학생들은 중학교에서 공통과학을 배우기 때문에 전체 학생들이 이러한 문항에 대하여 답하지 않았다. 이러한 변수들은 분석에서 삭제하였다. 학생 ID, 가중치, 검사일, 검사관 직책, 검사 언어 등의 변수 또한 학생의 수학 성취수준과 직접 연관이 없거나 우리나라 맥락에서 불필요한 변수이므로 분석에서 삭제하였다.

TIMSS는 전체 영역 및 각 세부 영역의 PV(plausible values)에 대응되는 수학 성취수준 또한 제공한다. 성취수준 변수는 PV 변수로부터 직접 계산된 값이므로 이 두 종류의 변수를 한 모형에 투입시키는 경우 PV 변수가 모형에서 인위적으로 큰 설명력을 가진다는 문제가 발생한다. 따라서 본 연구에서 수학 전체 영역에서의 성취수준 변수 5개(BSMIBM01 ~ BSMIBM05)만을 반응변수에 이용하고, 다른 성취수준 변수와 PV 변수들은 분석에서 제외하였다. 마찬가지로 개개의 설문변수는 모형에 포함시키되, TIMSS에서 설문 변수를 조합하여 자체 생성한 변수들(예: BSBGEML, BSDGEDUP 등)은 분석에서 제외하였다. 마지막으로 학생들의 각 변수에 대한 무응답 및 '잘 모르겠다' 반응 삭제 후 자료는 105개 변수(설명변수 100개, 성취수준 변수 5개)에 대한 2426명으로 정리되었다.

2. 반응변수와 설명변수

본 연구의 반응변수를 도출하기 위하여 5개의 성취수준 변수를 하나의 변수로 정리하였다. 이때 앙상블(ensemble) 방법에서 주로 이용되는 다수결 법칙(majority rule)을 이용하였다(Breiman, 1996). 즉, 어떤 학생의 성취수준이 1, 2, 1, 1, 2라면 다수결 법칙에 의하여 이 학생의 성취수준은 1이 된다. 다수결 법칙 후 우리나라 중학교 2학년의 수학 성취수준별 수와 비율은 <Table 1>과 같다.

<Table 1> TIMSS 2011 Student Levels after Majority Vote

Achievement Level	1	2	3	4	5	NA
Number of students (%)	25 (1.0%)	115 (4.7%)	342 (14.1%)	686 (28.3%)	1252 (51.6%)	6 (0.2%)

<Table 1>에서 NA인 6명은 성취수준이 예를 들어 1, 1, 2, 2, 3과 같이 어느 한 성취수준으로 결정하기 어려운 경우였다. 우리나라 중학교 2학년의 수학 성취수준은 가장 높은 수준인 '5'가 절반을 넘는 약 52%였고, 다음으로 높은 수준인 '4'가 28%, '3'이 14%, '2'와 '1'은 각각 약 5%와 1%에 불과하였다. 따라서 수학 성취수준을 '5'인 학생과 '5'가 아닌 학생으로 나누어 분석하는 것이 의미가 있을 것이라고 판단하였다.

총 100개의 설명변수 중 '예', '아니오'의 이분변수는 모두 더미코딩으로 변환하였고, '매우 동의한다'부터 '전혀 동의하지 않는다'의 리커트식 척도로 응답한 문항은 연속형 변수로 취급하여 분석하였다. 아버지와 어머니의 학력, 본인의 기대 교육수준 변수의 경우에도 연속형 변수로 분석하였다.

3. 분석모형과 소프트웨어

수학 성취수준이 '5'인 학생은 전체의 약 52%, 나머지 학생은 48%였다. 반응변수가 수학 성취수준이 '5'냐 아니냐는 것이었으므로 식 (2)와 같은 로지스틱 회귀모형을 분석모형으로 이용하였다.

$$\log\frac{P(G=1|X=x)}{P(G=0|X=x)}=\beta_0+\beta^T X \qquad (2)$$

LASSO는 변수의 척도에 민감하게 반응하므로, 코딩이 끝난 모든 변수를 표준화 후 식 (3)을

이용하여 추정하였다(Hastie et al., 2009).

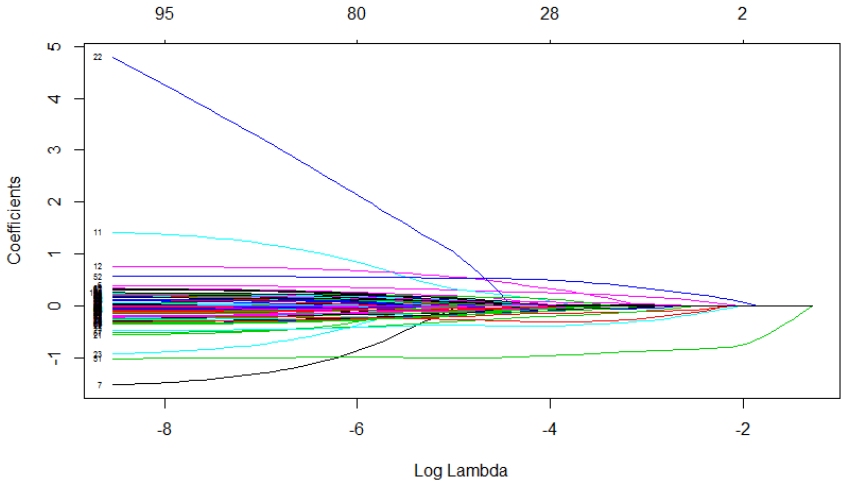
$$\max_{\beta}\left\{\sum_{i=1}^{N}[(y_i(\beta_0+\beta^T X_i)-\log(1+e^{\beta_0+\beta^T X_i})]-\lambda\sum_{j=1}^{P}|\beta_j|\right\} \quad (3)$$

본 연구에서는 자료를 7:3의 비율로 훈련자료와 시험자료로 나누고, 훈련자료로 모형을 구축한 후 시험자료로 모형을 평가하였다. 먼저 LASSO를 위하여 R의 glmnet library를 이용하였고, 교차타당화 시 cv.glmnet의 10-fold CV로 자료를 분석하였다. 로지스틱 회귀모형에 대한 교차타당화 시 이탈도(deviance)를 이용하여 시험자료에서의 정분류율(accuracy), 민감도(sensitivity), 특이도(specificity)와 비교하여 예측오차를 얻었다.

Ⅴ. 연구결과

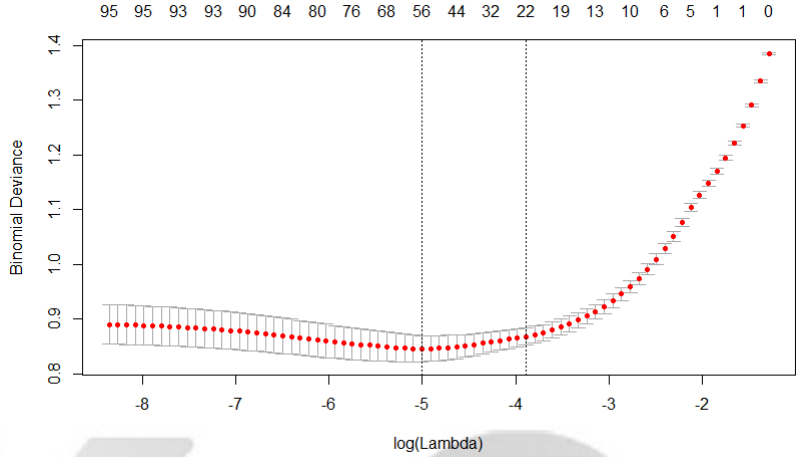
1. LASSO 모형 선택

[Figure 1]은 TIMSS 2011 한국 중학생 자료에 대한 LASSO 계수 프로파일이다. 각각의 선들은 100개 설명변수의 회귀계수가 벌점모수(λ) 값이 증가함에 따라 0으로 축소되는 것을 보여준다. 벌점모수(λ)가 커지면서 LASSO 계수가 0에 수렴하는 것을 확인할 수 있다.



[Figure 1] LASSO coefficients profile on TIMSS 2011 Korea data

[Figure 2]는 이탈도 기준 LASSO 축소추정 양상을 그래프로 보여준다. 두 개의 점선은 '1-standard error rule'(1-표준오차 법칙)을 이용하여 이탈도가 가장 작은 λ값으로부터 1-표준오차에 해당되는 지점을 표시한다.



[Figure 2] Cross-validation result with deviance

이탈도를 기준으로 1-표준오차를 이용하여 얻은 벌점모수 값은 0.02046009였다. 훈련자료의 정분류율, 민감도, 특이도는 각각 80.93%, 81.54%, 80.37%였고, 시험자료의 정분류율, 민감도, 특이도 또한 각각 79.34%, 78.46%, 80.29%로 거의 비슷하였다. 즉, 본 연구의 LASSO 모형은 일반화가능성이 상당히 높은 것으로 보인다.

2. LASSO 계수 해석

이탈도를 기준으로 한 0이 아닌 22개 LASSO 계수는 <Table 2>와 같다. 추정치의 분산을 줄이는 대신 편향이 늘어나는 방법인 LASSO는 정확한 편향을 측정하기 힘들기 때문에 추정치의 표준오차를 제시하지 않는다(Goeman, Meijer, & Chaturvedi, 2016; Yoo, 2016). 연구 결과, 100개 변수에서 선택된 22개 변수는 가정배경 변수 6개, 수학 관련 변수 12개, 그리고 과학 관련 변수 4개 변수로 정리되었다.

가정배경변인으로 가정의 장서 보유 수(BSBG04)와 인터넷 연결 유무(BSBG05E), 가정의 백과사전 보유 유무(BSBG05F), 어머니 교육수준(BSBG06A), 아버지 교육수준(BSBG06B), 그리고 부모와의 학교생활에 대한 대화정도(BSBG11B)가 모형에 포함되었다. 가정의 교육자원이 많을수록, 즉 집에 책이 많고 인터넷이 연결되며 백과사전을 보유할수록, 부모의 교육수준이 높을수록, 가정에서 부모와의 대화가 많을수록 학생의 수학성취수준이 5일 확률이 더 높았다.

수학 관련 12개 변수는 수학 태도 4문항, 수학적 자기효능감 7문항, 그리고 수학숙제 1문항으로 구성된다. 먼저 수학 태도와 자기효능감의 경우, 학생이 수학을 배우는 것을 좋아할수록, 수학에 자신이 있다고 생각할수록 성취수준이 5일 확률이 높았다. 더 자세하게 설명하면 다음과 같다. 수학 학습을 좋아한다(BSBM14A), 수학에서 재미있는 것을 많이 배운다(BSBM14D), 수학을 배우는 것을 좋아한다(BSBM14E), 수학을 잘 하는 것이 중요하다(BSBM14F)고 답할수록 성취수준이 5일 확률이 높았다. 총 14개 문항으로 구성된 수학적 자기효능감 문항 중 수학을 대체로 잘 한다(BSBM16A), 다른 학생들에 비해 수학이 쉽다(BSBM16B), 수학을 빨리 배운다(BSBM16D), 어려운 수학문제 풀기를 잘 한다(BSBM16F), 교사가 내가 수학을 잘 한다고 생각한다(BSBM16G), 원하는 대학에 들어가기 위해 수학이 필요하다(BSBM16L), 원하는 직업을 얻기 위해 수학을 잘 할 필요가 있다(BSBM16M)고 답할수록 성취수준이 5일 확률이 더 높았다.

수학과 과학 숙제 관련 문항이 각각 한 문항씩 모형에 포함되었다. 수학숙제 시간이 짧을수록(BSBM20B), 교사가 과학숙제를 적게 내줄수록(BSBS21A) 수학 성취수준이 5일 확률이 더 높았다. 다른 과학 관련 문항은 과학 태도에 관한 것으로, 과학 학습을 좋아한다(BSBS17E), 대체로 과학을 잘 한다(BSBS19A), 다른 학생에 비해 과학이 쉽다(BSBS19B)고 답할수록 수학 성취수준이 5일 확률이 더 높았다.

<Table 2> LASSO coefficients by deviance

Variable	Items and Labels	β
BSBG04	About how many books are there in your home? 1: '0-10 books' to 5: 'More than 200 books'	0.2442
BSBG05E	Do you have internet connection at your home? 1: 'Yes', 0: 'No' (recoded)	0.0495
BSBG05F	Do you have encyclopedia(a book or a CD) at your home? 1: 'Yes', 0: 'No' (recoded)	0.2970
BSBG06A	What is the highest level of education completed by your mother (or stepmother or female legal guardian)? 1: ISCED 1 or 2 or no school, 2: ISCED 2, 3: ISCED 3, 4: ISCED 4, 5: ISCED 5B, 6: ISCED 5A or First Degree 7: Beyond {ISCED 5A or First Degree}	0.0677
BSBG06B	What is the highest level of education completed by your father (or stepfather or male legal guardian)? The same labels as item BSBG06A	0.0250
BSBG11B	I talk about my schoolwork with my parents 1: 'Everyday or almost everyday' to 4: 'Never or almost never'	-0.0332
BSBM14A	I enjoy learning mathematics	-0.0888
BSBM14D	I learn many interesting things in mathematics	-0.0601

Variable	Items and Labels	β
BSBM14E	I like mathematics	-0.0188
BSBM14F	It is important to do well in mathematics	-0.0468
BSBM16A	I usually do well in mathematics	-0.9522
BSBM16B	Mathematics is more difficult for me than for many of my classmates	0.4832
BSBM16D	I learn things quickly in mathematics	-0.0249
BSBM16F	I am good at working out difficult mathematics problems	-0.1409
BSBM16G	My teacher thinks I can do well in mathematics classes with difficult materials	-0.1977
BSBM16L	I need to do well in mathematics to get into the college or university of my choice	-0.3078
BSBM16M	I need to do well in mathematics to get the job I want	-0.0353
BSBS17E	I learn many interesting things in science	-0.0974
BSBS19A	I usually do well in science	-0.3832
BSBS19B	Science is more difficult for me than for many of my classmates	0.0039
BSBM20B	About how many minutes do you usually spend on your mathematics homework? 1: 'My teacher never gives me homework in mathematics' 2: '1-15 minutes' to 5: 'more than 90 minutes'	-0.0084
BSBS21A	How often does your teacher give you homework in science? 1: 'Everyday' to 5: 'Never'	0.1125

Note: Items are presented in the order TIMSS provides.
When unspecified, items' labels are 1: 'Agree a lot' to 4: 'Disagree a lot'.

Ⅵ. 논 의

1. 기존 연구에서의 변수

기존 연구모형에서도 포함되었던 학생의 수학적 자기효능감, 수학에 대한 태도, 어머니의 교육수준, 그리고 가정 보유 장서 수와 같은 가정의 교육자원 변수가 학생의 수학 성취수준에 영향을 미치는 것으로 나타났다. 즉, 학생의 수학적 자기효능감, 태도, 그리고 어머니의 교육수준이 높고 (또는 좋고) 가정 교육자원이 많을수록 중학생의 수학성취도가 높았으며, 이는 기존 연구 결과와 일치한다. 본 연구에서는 아버지의 교육수준도 모형에 포함되었으며, SES의 간접적 측정 시 이용되었던 여러 가정 보유 품목 중 인터넷과 백과사전만이 선택된 점 또한 특기할 만하다. 그러나 학생의 나이와 성별, 가정에서 쓰는 언어, 학교 일에 대한 부모 관여, 학교에서의 컴퓨터 이용, 학교폭력 및 학교왕따와 같은 기존 연구 변수들은 LASSO 모형에서 선택되지 못했다.

본 연구모형에서 수학숙제 시간이 짧을수록 학생의 수학성취도가 높았다. 중학생의 수학숙제 시간과 성취도 간 관계는 TIMSS-R(TIMSS 1999) 연구에서 자주 이용된 적 있으나, 이후 두 건의 TIMSS 2003 연구 외에는 찾아보기 힘든 변수다. 따라서 숙제 관련 변수는 2007년과 2011년 자료에 초점을 맞춘 본 연구의 선행연구에서 다뤄지지 않았다. 중학생의 수학숙제 시간과 수학성취도 간 관계를 분석한 TIMSS 20003 연구 결과는 정적(Zhu & Leung, 2012)이거나 관계 없음(Mikk, 2006)으로 일관되지 않았다. 교과와 자료는 다르지만, 오스트리아, 네덜란드, 일본, 한국, 미국 중등학생의 PISA 읽기 자료를 분석한 연구에서는 숙제 시간과 성취도 간 전반적으로 부적인 관계가 있었다(Dettmers, Trautwein & Ludtke, 2009; Trautwein, 2007). Epstein & Van Voorhis(2001)은 숙제시간과 학업성취도 간 관계는 단순하지 않으며, 연구에 따라 상반된 결과가 도출된다고 요약하였다. TIMSS 학생설문지가 학생들의 숙제 시간과 교사의 숙제 부여 빈도만을 다루는 반면, TIMSS 교사설문지는 숙제를 수업에서 어떻게 이용하는지에 대해서도 자료를 수집한다. 후속 연구에서 TIMSS 교사설문지까지 포괄하여 숙제와 학업성취도 간 심층적인 분석이 수행될 필요가 있다.

2. 본 연구에서 새롭게 파악된 변수

기계학습 기법을 적용한 본 연구를 통해 기존 연구에서 자주 이용되었던 변수뿐만 아니라, 기존 연구에서 고려하지 못했던 변수들을 새롭게 파악할 수 있었다. 이러한 변수로 과학숙제 관련 변수 및 과학 태도 및 과학적 자기효능감 문항들이 있었다. 연구 결과, 교사가 과학숙제를 적게 내줄수록, 학생의 과학적 자기효능감 및 과학에 대한 태도가 높을수록(또는 긍정적일수록) 수학성취도가 더 높았다. 주목할 점으로, 학생의 과학적 자기효능감 및 과학숙제 부여 빈도와 같은 과학 관련 문항은 그동안 TIMSS 수학성취도 모형에서는 이용되지 않았던 변수들이다.

TIMSS는 교사 설문지로 인하여 수학과 과학 교과의 교수학습 전반에 대한 변수들을 학생의 학업성취도 설명변수로 이용할 수 있는 장점이 있음에도 지금까지 성취도와 숙제 관련 변수에 대한 TIMSS 연구는 관례적으로 같은 교과로 한정되어 왔다. 자료를 TIMSS만으로 한정짓지 않고 선행연구들을 분석할 때에도 교과를 달리한 성취도 연구는 찾기 힘들었다. 직접적인 성취도 연구는 아니지만, 수학적 자기효능감과 과학관련 진로선택 관계를 다룬 Lent, Lopez, & Bieschke (1991)의 연구를 찾을 수 있었다. 반면, 과학적 자기효능감과 수학관련 진로라든지 과학적 자기효능감과 수학 성취도를 다룬 연구는 없는 것으로 보인다.

우리나라 중학생의 경우 수학성취도와 과학에 대한 태도 및 과학적 자기효능감이 정적인 관계가 있었고, 과학숙제 부여 빈도와 수학성취도는 부적인 관계가 있었다. 과학에 대한 태도 및 과학적 자기효능감과 수학성취도 간 정적인 관계는 수학과 과학의 학문적·실제적 연관성에서 기인한

것으로 사료된다(Lent et al., 1991). 반면, 과학숙제 부여 빈도와 수학성취도 간 부적인 관계는 과학 교과 관련 변수와 수학성취도에 대한 기존 TIMSS 연구가 전무한 상황에서 해석하기 쉽지 않다. 이를 주제로 한 후속 연구가 필요하다.

3. 연구 의의

그간 TIMSS 연구는 소수의 선택된 변수를 HLM, SEM 등의 모형에 투입하여 학생의 수학성취도에 영향을 미치는 요인을 연구해 왔다. 그러나 이러한 회귀모형으로는 몇 천 명의 응답자가 있다고 하더라도 몇 백 개의 변수를 모두 투입한 모형은 수렴의 문제로 인하여 모형화가 어려우며, 모형화된다 하더라도 과적합 문제 또한 심각할 수 있다(Yoo, 2016). 본 연구는 기계학습적 접근법인 LASSO를 통하여 TIMSS가 제공하는 수백 개의 변수를 모두 분석에서 고려하였다. 자료 정리 후 모형구축에 이용된 설명변수 100개 중 22개 변수가 선택되었으며, 모형의 예측정확도는 79.34%였다. 다시 말해, 모형구축에 이용되지 않은 새로운 학생들의 수학성취도를 예측할 때 본 연구모형은 약 80%의 상당히 높은 정확도를 보였다.

요약하자면, 본 연구는 학업성취도 예측을 위하여 교육 관련 대용량 자료에 기계학습 기법을 처음으로 적용한 연구로 의의가 있다. 순수하게 자료만을 이용하는 기계학습 기법을 적용함으로써 이론을 중시하는 기존 연구에서 고려하지 못했던 새로운 변수들을 파악할 수 있었다. 이렇게 새롭게 탐색된 변수들을 후속 연구에서 모형화함으로써 모형의 설명력을 높임과 동시에 연구 지평 또한 넓힐 수 있을 것이다.

References

Azina, I. N., & Halimah, A. (2012). Student factors and mathematics achievement: Evidence from TIMSS 2007. *Eurasia Journal of Mathematics, Science & Technology Education, 8*, 249-255.

Breiman, L. (1996). *Bagging predictors*. Machine Learning, 26, 123-140.

Cheng, Q. (2014). Quality mathematics instructional practices contributing to student achievements in five high-achieving Asian education systems: An analysis using TIMSS 2011 data. *Frontiers of Education in China, 9*, 493-518.

Dettmers, S., Trautwein, U., & Ludtke, O. (2009). The relationship between homework time and achievement is not universal: Evidence from multilevel analyses in 40 countries. *School Effectiveness and School Improvement, 20*, 375-405.

Engel, L. C., Rutkowski, D., & Rutkowski, L. (2009). The harder side of globalisation: Violent conflict and academic achievement. *Globalisation, Societies and Education, 7*, 433-456.

Epstein, J. L., & Van Voorhis, F. L. (2001). More than minutes: Teachers' roles in designing homework. *Educational Psychologist, 36*, 181-193.

Goeman, J., Meijer, R., & Chaturvedi, N. (2016). *L1 and L2 Penalized Regression Models*. Retrieved May 30, 2016 from https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction (2nd ed.)*. New York: Springer.

Kareshki, H., & Hajinezhad, Z. (2014). A multilevel analysis of the role of school quality and family background on students' mathematics achievement in the Middle East. *Universal Journal of Educational Research, 2*, 593-602.

Kim, W. (2012). A study on factors of education's outcome in liberal computer education using regression and data mining analysis. *Korean Journal of General Education, 6*, 743-767.

Kim, W. (2013). A study on factors of the academic achievement in computer training courses as the liberal arts in university. *Journal of the Korean Association of Information Education, 17*, 433-447.

Kim, J., & Yong, H. S. (2014). An analysis of specialized vocational high school's educational outcome using data mining technique. *Journal of Korean Association of Computer Education, 17*(6), 21-33.

Lent, R. W., Lopez, F. G, & Bieschke, K. J. (1991). Mathematics self-efficacy: Sources and relation to science-based career choice. *Journal of Counseling Psychology, 38*,

424-430.

Matsuoka, R. (2014). Disparities between schools in Japanese compulsory education: Analyses of a cohort using TIMSS 2007 and 2011. *Educational Studies in Japan: International Yearbook, 8*, 77-92.

Mikk, J. (2006). *Students' homework and TIMSS 2003 mathematics results*. ERIC Clearinghouse. (ERIC No. ED491866)

Mills, J. D., & Holloway, C. E. (2013). The development of statistical literacy skills in the eighth grade: exploring the TIMSS data to evaluate student achievement and teacher characteristics in the United States. *Educational Research and Evaluation, 19*, 323-345.

Ng, K. T., Lay, Y. F., Areepattamannil, S., Treagust, D. F., & Chandrasegaran, A. L. (2012). Relationship between affect and achievement in science and mathematics in Malaysia and Singapore. *Research in Science & Technological Education, 30*, 225-237.

Silver, D., Huang, A., Maddison, C., Guez, A., Sifre, L. van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*, 484-489.

Sulku, S. N., & Abdioglu, Z. (2015). Public and private school distinction, regional development differences, and other factors influencing the success of primary school students in Turkey. *Educational Sciences: Theory & Practice, 15*, 419-431.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of Royal Statistics Society B, 58*, 267-288.

Trautwein, U. (2007). The homework-achievement relation reconsidered: Differentiating homework time, homework frequency, and homework effort. *Learning and Instruction, 17*, 372-388.

Tsai, L. T., & Yang, C. C. (2015). Hierarchical effects of school classroom, and student-level factors on the science performance of eighth-grade Taiwanese students. *International Journal of Science Education, 37*, 1166-1181.

Winnaar, L. D., Frempong, G., & Blignaut, R. (2015). Understanding school effects in South Africa using multilevel analysis: Finding from TIMSS 2011. Electronic *Journal of Research in Educational Psychology, 13*, 151-170.

Yoo, J. E. (2016). An analysis case of educational panel data through a data mining technique: A penalized regression with KYPS data. *Asian Journal of Education, 17*, 1-19.

Zhu, Y., & Leung, F. K. S. (2012). Homework and mathematics achievement in Hong Kong: Evidence from the TIMSS 2003. *International Journal of Science and Mathematics Education, 10*, 907-925.

Abstract

TIMSS 2011 Predictors Relating to Korean 8th Graders' Mathematics Achievement, Explored Via Machine Learning

Yoo, Jin Eun (Korea National University of Education)

A substantial body of research has been conducted on factors relating to students' math achievement with TIMSS. However, most studies have focused on selected a few factors instead of utilizing hundreds of variables TIMSS provides, and have employed conventional statistical methods. This study aimed to investigate possible sets of predictors from a totally different approach: LASSO, currently one of the most popular machine learning techniques. Korean 8th graders' TIMSS 2011 were used as the sample, and the prediction accuracy of the LASSO model was about 80% with the selected 22 out of 100 predictors. As results, students' math efficacy, attitudes toward math, mother's education level, and home educational resources including amount of books at home were influential to their math achievement, which was consistent with previous studies. Additionally, math homework completion time, student's science self-efficacy, and science homework frequency were newly found important predictors. Implications and future research topics are discussed.

Key words : Machine learning, LASSO, Penalized regression, TIMSS, Math achievement