Journal of the Korea Academia-Industrial cooperation Society
Vol. 24, No. 11 pp. 718-725, 2023

https://doi.org/10.5762/KAIS.2023.24.11.718
ISSN 1975-4701 / eISSN 2288-4688

Computer-Based Test에서 반응시간과 인공지능기법을 이용한 학생 시험행동의 분류

이미정[1,2], 채유미[2], 박석건[3], 박일용[4*]
[1]단국대학교 의과대학 소아청소년과학교실, [2]단국대학교 의과대학 의학교육학교실,
[3]단국대학교 명예교수,[4]단국대학교 의과대학 의공학교실

Grouping of Student Examination Behaviors using Response Time and Artificial Intelligence Technique in Computer-Based Test

Mee Jeong Lee[1,2], Yoo Mi Chae[2], Seokgun Park[3], Ilyong Park[4*]
[1]Department of Pediatrics, Dankook University College of Medicine, Cheonan, Korea
[2]Department of Pediatrics, Dankook University College of Medicine, Cheonan, Korea
[3]Emeritus Professor, Dankook University, Cheonan Korea
[4]Department of Biomedical Engineering, Dankook University College of Medicine, Cheonan, Korea

요 약 Computer-based test (CBT)를 이용한 평가에서 기존 지필시험에서는 확인할 수 없었던 항목을 도출하여 학생의 시험행동을 분석하였다. 연구 대상은 D의대 의학과 3학년 학생 40명으로, CBT로 한 과목의 총괄 시험을 시행하고, 데이터 분석을 통해 학생 시험행동 유형을 분류하였다. 각 문항 당 반응시간(response time, RT)과 전체 문항에 대한 총반응시간(총RT)를 정의한 후, 성적과 함께 백분위로 나타내어 2차원 데이터를 얻었다. K-means 클러스터링 및 인공신경망을 활용하여 시험행동 유형을 4개의 그룹으로 분류하였다: 높은 성적/짧은 총RT(1그룹), 높은 성적/긴 총RT(2그룹), 낮은 성적/긴 총RT(3그룹), 낮은 성적/짧은 총RT(4그룹). 1그룹은 뛰어난 학생들로 예상되었고, 2그룹은 바람직한 시험행동을 가진 학생으로 판단하였다. 3그룹은 학습 능력과 방법에 대한 개선이 필요한 그룹으로 분류되었다. 4그룹은 빠른 응답을 보이지만 시험 문항이 요구하는 지식과 학습 목표를 이루지 못하는 경향을 보여, 보충 학습과 신중한 시험 태도에 대한 지도가 필요하다고 판단하였다. 이러한 결과는 CBT에서 RT를 측정하여 성적과 결합함으로써 학생들의 시험행동을 분류할 수 있음을 보여주며, 이는 학생의 개인별 맞춤형 학습 지도 및 상담에 유용하게 활용될 기초자료가 될 것이다.

Abstract In evaluations using the computer-based testing (CBT) system, student examination behavior can be classified by deriving information that could not be observable in paper-based tests. This study focused on 40 fifth-grade students at D Medical College, conducting a final examination for one subject using CBT. Examination behaviors were categorized by defining response time (RT) for each question and total response time (TRT) for the entire examination. These were presented as percentiles along with student grades, resulting in two-dimensional data. Examination behaviors were classified into 4 groups by using K-means clustering and artificial neural networks: high-grade/short TRT (Group 1), high-grade/long TRT (Group 2), low-grade/long TRT (Group 3), and low-grade/short TRT (Group 4). Group 1 students were anticipated to be excellent, while Group 2 exhibited desirable examination behavior. Group 3 was evaluated as needing improvement in learning abilities and strategies. Group 4 showed a quick response but struggled to meet the knowledge and learning objectives of the examination, suggesting a requirement for supplementary learning and guidance on a careful test approach. These findings demonstrate that by incorporating RT measurements into CBT and combining them with grades, student examination behaviors can be classified, offering the potential for personalized learning guidance and counseling.

Keywords : High-stake Test, Examination Strategy, K-means Clustering, Machine Learning, Student Counseling

*Corresponding Author : Ilyong Park(Dankook Univ.)
email: piyong@dankook.ac.kr
Received October 4, 2023 Revised November 2, 2023
Accepted November 3, 2023 Published November 30, 2023

1. 서론

시험의 도구로 컴퓨터를 이용하는 시험 방법(Computer Based Test, 이하 CBT)은 종이와 연필을 이용하는 시험에 비해 여러 가지 장점을 가지고 있다. 우선 종이에 인쇄하는 지필시험과 달리 사진이나 소리, 색채 등을 자유롭게 사용할 수 있고, 멀티미디어를 사용하여 시험문제를 실제 상황에 가깝게 만들 수도 있다[1]. 또 다른 장점은 빠른 채점과 시험 결과의 분석이다. 시험을 끝내자마자 바로 채점해서 피드백을 할 수 있는 장점이 있다[1]. CBT 시스템에 문항분석을 할 수 있는 서브시스템을 같이 운용하면 채점 결과를 다시 입력해서 계산하는 절차를 거치지 않고도 바로 문항분석을 해서 출제자에게 피드백을 해줄 수 있다[1]. 난이도, 변별도 등 분석이 완료된 문제들을 바로 데이터베이스에 연결해서 문제은행을 구축하도록 할 수도 있다[2].

Yim[3]은 컴퓨터 기반 보건의료인 국가시험 개발연구 보고서에서 CBT를 이용하여 장기적으로는 상시시험체제로 전환하거나, 문항 수를 줄인 컴퓨터화 능력 적응검사(Computerized Adaptive Test, 이하 CAT) 도입을 고려해 볼 수 있다는 제안을 한 바 있다. 또한 우리나라도 전문직 시험에 CAT를 도입하는 것에 대해 진지하게 고민한 예[4]도 있었다.

이러한 장점들 외에, CBT는 연필과 종이를 쓰는 방법으로는 지금까지 전혀 불가능했던 부가적인 가치를 제공할 수 있다. 즉, 시험을 보는 중에 학생이 문항마다 소비하는 시간을 실시간으로 기록할 수 있고, 이 결과를 이용해서 학생이 시험 중에 하는 행동(또는 학생이 시험을 보면서 사용하는 시험전략)을 추리해 볼 수 있다[3,5,6].

학생이 컴퓨터에서 해당 시험문제를 열어서 답을 입력할 때까지 사용한 시간을 반응시간(response time, 이하 RT)이라고 하면, 어려운 문제에서는 긴 RT를, 쉬운 문제에서는 짧은 RT를 보여주고, 실력이 없는 학생은 긴 RT를, 실력이 있는 학생은 짧은 RT의 시험 결과와 연결된다[7]. 다만 관계가 이렇게 단순하지만은 않아서 문제가 너무 어렵거나, 아예 문제를 풀 의지가 없이 아무 답이나 찍는(rapid-guessing behavior) 경우도 짧은 RT를 보일 수 있다. Wise[8]는 시험 결과를 학생의 개인성적에는 반영하지 않아서 학생이 진지하게 문제를 풀 동기가 없을 수도 있는 학력 측정 시험 등에서 RT를 측정하면, 학생들이 진지하게 문제를 풀었는지, 아니면 그냥 빠르게 아무 답이나 입력했는지를 알 수 있다고 보고하였다.

저자들은 위에서 기술한 것과 같이 RT가 학력 측정 시험 또는 저부담(low-stake) 시험에서 시험문제의 질을 나타내는 지표가 될 수 있을 뿐만 아니라, RT를 시험 성적, 시험의 난이도 등과 적절하게 결합하면 학생들이 시험 중에 하는 시험행동 또는 시험에 임하는 전략을 들여다볼 수 있는 지표가 될 수도 있으리라는 가설을 세우고, 여기서 일정한 패턴들이 나타나는지를 보고자 하였다.

고부담(high-stake) 과목에서 시험 성적 결과와 CBT에서 얻은 RT를 이용하여 학생들의 시험행동을 합리적인 설명이 가능한 몇 개의 패턴으로 분류할 수 있다면, RT는 학생들의 시험행동을 판단할 수 있는 정량적인 근거 자료의 하나가 될 것이다. 본 연구에서는 총RT가 학생들의 시험행동을 탐색하는 지표로 이용될 수 있음을 보이고자 하였다. 또한 그 결과를 개별 학생의 학습 특성에 맞춘 맞춤형 학업 지도에도 유용하게 사용할 수 있을 것이다. 따라서 본 연구는 의과대학생의 CBT 시험행동 분석을 통해 학습 특성별 맞춤형 학업 지도에 활용할 기초자료가 될 것이다.

2. 본론

2.1 연구 대상

2019년 2월에 시행한 의학과 3학년 소아과학 기말고사(CBT) 결과를 사용하였다. 시험 문항은 총 92개이었으며, 모두 객관식 문제였다. 응시한 인원은 총 44명이었다. 이 중에서 한국어가 모국어가 아닌 외국인 학생 4명을 배제하고 최종적으로 40명의 시험 결과를 이용하였다(IRB No. DKUH 2021-11-008).

2.2 연구 방법

2.2.1 RT 측정

학생이 CBT에 로그인하여 문제를 선택한 후 해당 문제의 답을 입력할 때까지의 시간을 문제당 RT로 정의하였다. 제시된 모든 문제에서 측정된 RT를 합한 결과를 총RT로 하였다. 시험문제에 로그인하여 시험을 끝내고 로그아웃할 때까지 걸린 전체 시간은 총소요시간으로 하였다. RT의 정의는 시험을 시작하기 전에 미리 CBT 시스템에 프로그램해 두었다. CBT 시스템용 서버 소프트웨어는 Apache, MySQL, PHP 기반으로 동작하는 TCExam (ver. 11)을 사용하였다.

2.2.2 RT와 난이도의 비교

연구에 이용한 시험문제의 적절성을 확인하기 위해 고전검사이론에 따른 시험문제의 난이도와 변별도를 계산하였으며, 난도가 높아짐에 따라 RT의 변화가 있는지를 확인하기 위하여 난이도와 RT의 관계를 비교하였다.

2.2.3 인공지능기법(artificial intelligence technique, AI 기법)을 이용한 분류

학생의 시험행동과 관련 있는 학생 성적백분위와 총RT백분위로 이루어진 2차원 데이터에 대해 K-means 클러스터링(clustering) 및 인공신경망을 적용하여 유형을 분류하였다. K-means 클러스터링 및 인공신경망의 구현을 위해 LabVIEW 2015 및 Machine Learning Toolkit (National Instruments, US)를 사용하였고, 사용된 인공신경망은 낮은 컴퓨터 사양에서도 빠른 분류 속도를 보일 수 있도록 단순한 fully-connected layer 구조를 채택하였다.

최고 점수를 1.0으로 최저 점수를 0점으로 하여 각 학생의 점수를 백분위로 재정리하고, 총RT도 가장 짧은 시간을 0, 가장 긴 시간을 1.0으로 설정하여 백분위로 환산하였다.

이렇게 백분위로 환산한 각 학생의 성적과 총RT를 각각 X축과 Y축으로 결정하여 2차원 데이터로 정리하였다. 2차원 데이터를 K-means 클러스터링 및 인공신경망을 적용하여 학생 시험행동 유형을 분류하도록 하였다. K-means 클러스터링은 비지도 기계학습 알고리즘의 하나로서, 여러 데이터 포인트 중에서 유사한 특성을 가진 데이터들을 서로 다른 그룹들로 분류하여 묶어주는 클러스터링 기술이다[9]. 사전에 정해진 K개의 클러스터로 분류하여 나누는 방식으로 작동하며, 각 클러스터는 중심점(centroid)이라 불리는 포인트로 대표되며, 이 중심점은 해당 클러스터 내의 모든 포인트와의 거리를 최소화하는 방식으로 결정된다. 본 연구에서 사용된 K-means 클러스터링 알고리즘의 첫 번째 단계인 초기화에서는 먼저 네 가지 학생 시험 수행 태도 분류 클러스터의 초기 중심들을 선택하였다. 이 초기 중심점은 성적백분위와 총RT백분위를 각각 X, Y축으로 하는 2차원 공간상에서 임의의 값으로 무작위 선택하였다. 그다음에 각 데이터 포인트와 설정된 K개의 중심점 간의 유클리드 거리들을 계산하여 가장 가까운 중심점으로 대표되는 클러스터에 해당 데이터 포인트를 할당하였다. 이렇게 4개의 클러스터에 할당된 데이터 포인트들을 평균한 것을 이용하여 새로운 중심점을 계산하였다. 이러한 클러스터 할당과 새로운 중심점 업데이트 과정을 계속 반복하여 클러스터 중심점이 더 이상 크게 변하지 않거나, 정해진 반복 횟수에 도달하면 K-means 클러스터링 알고리즘을 종료하였다.

3. 연구 결과

3.1 난이도, 변별도, 난이도와 총RT

3.1.1 고전문항분석

난이도와 변별도는 각각 Table 1, Table 2와 같았다. 매우 쉬움 또는 쉬움인 문제가 총 76개로 전체 시험 문제의 82.6%를 차지하여, 기초 능력 확보를 판단하기에는 좋은 시험으로 판단하였다. 어려움 또는 매우 어려움인 문제가 총 12문제로 전체의 13.0%를 차지하여 학생들의 역량 변별에도 바람직한 것으로 판단하였다.

Table 1. Difficulties of items

Difficulty	very easy	moderately easy	moderately difficult	very difficult	extremely difficult
Number of item	69	7	4	8	4
Mean of RT(sec)	31.7	40.9	55.9	58.7	54.1

Table 2. Discrimination index of items

Discrimination index	-0.1	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
Number of item	2	14	12	10	12	14	7	13	5	3

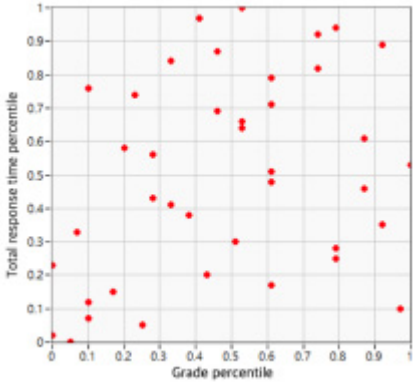
3.1.2 난이도와 RT

전체 학생의 한 문항 당 평균 RT는 36.8초이었다. 매우 쉬운 문제는 평균 31.7초, 쉬운 문제는 40.9초, 적절한 문제는 55.9초, 어려운 문제는 58.7초, 매우 어려운 문제는 54.1초였다.

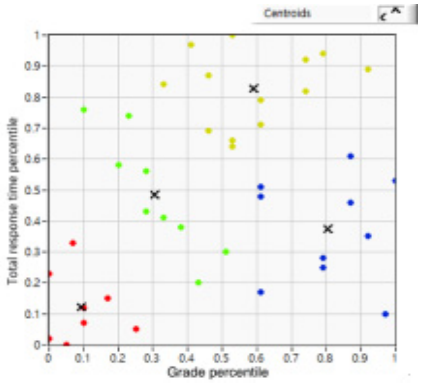
매우 어려운 문제 4개를 제외하면 문제 난이도에 따라 RT가 길게 나왔다. 매우 어려운 문제 중 모든 학생이 틀린 76번 문제는 RT가 평균 29초였고, 이 문제를 뺀 다른 3개 문제의 RT 평균은 62.4초이었으므로 전체적으로 RT와 난이도가 비례하였다.

3.2 성적백분위와 총RT백분위

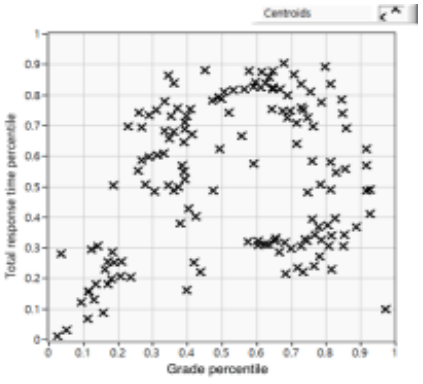
시험 결과의 성적백분위와 총RT백분위로 이루어진 입력 데이터는 Fig. 1(a)과 같다. 이 데이터에 대해 일종의 비지도 기계학습법인 K-means 클러스터링 기법을 이용하였고, Fig. 1(b)과 같이 4개의 응시자 시험 태도 유형 그룹으로 분류되었다. 이런 분류 기법은 데이터 분포 특성을 반영하여 알고리즘에 의해 선정된 분류 기준이 설정되는 장점을 확인할 수 있었다. 그런데, K-means 클러스터링 기법의 특성상 학습 초기 설정값에 따라 분류 결과가 달라질 수 있으므로, 같은 입력 데이터에 대해 K-means 클러스터링을 1,000번 반복 적용하여 Fig. 1(c)과 같이 각 분류 그룹의 중심점(centroid)들을 얻었다. 이를 통해 성적백분위와 총RT백분위 데이터의 2차원 공간에서의 각 분류 그룹의 중심점 분포를 확인하였다. Fig. 1(d)과 같이 1,000번의 반복된 분류 결과의 중심점 데이터에 대해 K-means 클러스터링을 실시하여 4개 유형 그룹으로 분류되었다. 이러한 그룹별 분류 분포 중심점 특성을 가진 인공신경망 기반의 분류 모델을 만들기 위해 그룹별 중심점 데이터를 Fig. 1(e)의 구조를 가지는 3층 구조의 인공신경망에 학습시켰다. 입력층은 2개의 노드, 중간 은닉층은 10개의 노드, 출력층은 4개의 노드로 구성하였다. Fig. 1(f)은 생성된 인공신경망 기반의 분류 모델을 이용한 교과목 시험 결과에 적용하여 학생 시험 수행 성향 분류 결과를 보여준다. 최종적으로, 평가자가 아닌 기계학습과 인공신경망을 이용하여 각 시험자를 총 4개의 그룹(높은 성적과 짧은 총RT(1그룹), 높은 성적과 긴 총RT(2그룹), 낮은 성적과 긴 총RT(3그룹), 낮은 성적과 짧은 총RT(4그룹))으로 분류하였다.



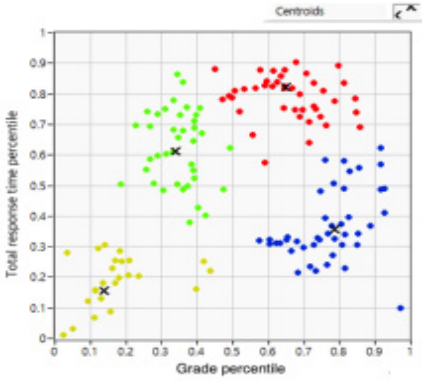
(a)



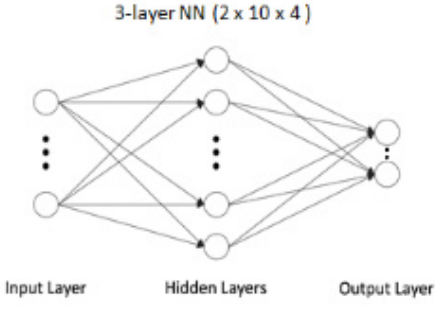
(b)



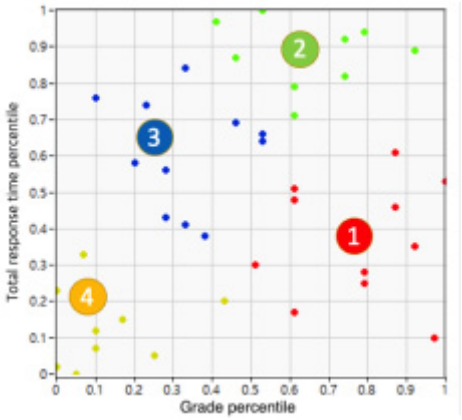
(c)



(d)



(e)



(f)

Fig. 1. Student examination behavior type classification results applied to subject A test results using the proposed K-means clustering based classification. (a) Input data, (b) an initial classification into 4 groups by K-means clustering (classification results vary depending on the initial input value), (c) distribution of all centroids of classification groups according to every K-means clustering iterations, (d) K-means clustering classification results for the centroid points, (e) artificial neural network structure to learn the centroid point classification results, and (f) final classification result using the learned artificial neural network.

4. 논의

학교는 학생의 성취도와 실력을 파악하기 위해서 시험을 부여한다. 학생은 시험을 통해 자신의 수준을 확인하며, 이것을 기준점으로 삼아 다음 단계로 배움을 확장하게 된다. 또한 강의자는 수강한 학생들의 성적으로 프로그램의 성공 여부를 판단하기도 한다.

시험은 필기시험, 실기시험, 구술시험, 현장평가 등 여러 가지 형태가 있을 수 있다. 이 중에서도 필기는 시험이 끝난 후에 객관적인 자료들을 수집하기가 쉬우며, 이를 분석하여 “시험문제가 학생의 수준을 평가하기에 적절했는가?”를 평가할 수 있다[4].

필기시험은 CBT 형태로 전환하기가 쉽다. CBT는 내용 면에서는 색채, 동영상, 소리 등을 사용하여 문제를 다채롭게 낼 수 있으며, 결과의 처리 면에서는 시험 후 빠르게 채점을 할 수 있고, 시험 결과 분석 또한 빠르게 내놓을 수 있다[1]. 분석 결과물을 응용하는 CAT도 가능하다[3]. 이렇게 필기시험의 내용을 풍부하게 해주고, 결과 처리를 아주 빠르게 해주며, 응용도 할 수 있다는 점에서 현재의 CBT는 증강된 필기시험이라고 할 수 있다.

그런데 CBT에서 추출한 RT는 ‘증강’에 그치지 않고 필기시험에서는 얻을 수 없는 전혀 새로운 정보를 제공한다. 즉, 시험을 끝낸 후의 결과를 분석의 대상으로 삼을 수밖에 없는 기존의 고전검사이론과 문항반응이론과는 달리, 학생이 시험 중에 하는 행동에 대한 단서를 제공할 수 있는 것이다. 학생이 CBT에서 시험문제를 열어서 답을 입력할 때까지의 시간을 측정함으로써 시험을 보는 동안의 학생의 사고 과정(시험행동)을 반영하는 지표가 될 수 있다.

국내에서는 Chae, Park과 Park[10]이 RT와 고전문항분석 결과를 비교하여 RT가 시험의 난이도와 비례하기는 하지만, 완전히 일치하지는 않는다는 결과를 보고한 바 있다. 어려운 문제를 깊이 생각해서 어떻게든 풀어보려는 학생과, 그냥 빨리 답을 찍고 다른 문제로 넘어가는 학생이 있다고 하면, 시험 후에 계산된 난이도는 같지만, 시험 중에 측정한 RT는 다르게 나올 것이다.

저자들은 RT와 난이도를 비교하였다. 그리고 RT와 성적을 결합하여 학생을 몇 개의 합리적인 설명이 가능한 그룹으로 분류할 수 있었다. 이 연구에서도 선행 연구와 마찬가지로 난이도가 높은 문제일수록 RT가 길어지는 비례관계가 나타났다.

RT와 성적을 결합하여 분류된 그룹은 [결과]와 같았다. 본 연구에서 분류 방법으로 사용된 K-means 클러스터링의 장점은 구현이 간단하고 빠르며, 본 연구의 대상인 의과대학 CBT 데이터와 같은 대용량 데이터 세트에도 적용할 수 있다는 점이다[11,12]. 하지만 초기 중심점의 선택이 클러스터링 결과에 영향을 미치며, 이상치(outlier)에 민감하게 작동할 수 있다[13]. 본 연구에서는 임의의 초기 중심점 선택에 따른 각 클러스터의 중심점들의 위치 변동성을 반영하기 위해, 같은 성적백분위와 총RT백분위 입력 데이터에 대해 총 1,000번의 K-means 클러스터링을 수행하여 총 4,000개의 클러스터 대표 중심점 중에서 중복되는 포인트는 제외하고 대표 중심점 포인트들을 대상으로 다시 4개의 그룹으로 K-means 클러스터링을 하였다. 그런 다음, 임의의 초기 중심점 설정값에 따라 4개 그룹의 클러스터링 중심점 변동성을 반영하여 분류된 4개의 클러스터 데이터 포인트들을 인공신경망으로 학습을 시켰다. 이렇게 학습된 인공신경망은 각 데이터 포인트가 성적백분위와 총RT백분위를 각각 X, Y축으로 하는 2차원 공간에서 어떤 군집 특성을 가지는지에 따라 별도의 K-means 클러스터링

과정 없이 미리 학습된 4개의 그룹 중 하나의 클러스터로 바로 분류될 수 있도록 해주는 장점이 가지게 된다. 또한, 각 데이터 포인트가 속하게 되는 그룹이 어느 정도의 확률로 그렇게 분류가 되었는지 정량적 척도를 제공하며, 이 척도는 특정 학생의 (성적백분위, 총RT백분위) 데이터 포인트가 분류된 시험 수행 성향 그룹의 대표 중심점 근처에 위치하는지, 다른 그룹과의 경계에 위치하는지를 정량적 수치로써 활용할 수 있다.

한편, 본 연구에서 제안한 기계학습 및 인공신경망을 이용한 학생 시험 수행 태도 분류 알고리즘은 하나의 학급 단위의 학생 성적백분위 및 총RT백분위 데이터 분포 특성을 기반으로 하므로 각 학급 내에서의 학생들의 상대적인 그룹 분류가 된다는 특징이 있다.

92문항에 2시간 이상의 충분한 시험 시간을 배정하여 역량 검사(pure power test)가 가능하게 하여[14], 시간이 부족하여 시험문제를 다 풀지 못하는 예는 없게 하였다. 소아과학은 시험 성적 하위 10-20% 학생들에게 재시를 주고, 재시에서도 충분한 성적을 보이지 않으면 “F” 학점을 준다. “F”가 나오면 유급이므로 소아과학은 고부담 과목이다. 따라서 학생들은 최대한의 역량을 발휘하여 시험에 임하였을 것이다.

RT는 CBT에서 측정할 수 있는 지표 중 하나로, 시험문제를 읽기 시작해서 답을 표시할 때까지의 시간으로 시험자가 ‘문제를 처리하는 데 사용한 시간’이 된다. 이 시간은 시험문제의 난이도뿐만 아니라 학생의 사고 과정 또는 시험행동(시험전략)을 반영하게 될 것이다. 측정된 RT가 길다면 문제 자체가 어렵거나, 혹은 그 학생의 능력이 부족하다고 볼 수 있다. 또 RT가 지나치게 짧다면, 문제가 매우 쉽거나, 문제를 제대로 풀지도 않고 답을 표시했거나, 과거에 경험하였던 문제라 답을 미리 알고 있었을 가능성이 있다. RT는 문제의 난도가 증가하면 증가한다. 그러나 시험자의 능력을 벗어난 난도를 보일 때 찍기 행태를 보여 RT가 감소하는 경향이 있다[10]. 총RT가 평균보다 과하게 감소한 학생은 찍기 행태를 주로 하여 본인의 역량을 충분히 발휘하지 않았을 수 있다. 충분한 시험 시간을 주었기에 시간에 쫓겨서 “찍기”를 한 것이 아니므로 학생의 시험 보는 태도에 문제가 있다고 해석할 수 있다. RT를 시험 성적과 결합하여 분석하면, 어떤 상황에 해당하는지 단서를 찾을 수 있을 것이다, RT를 다른 요인들과 조합하여 측정하는 것도 가능하다. 예를 들면, 시험문제를 다시 점검할 때, 기존의 답이 맞았을 때, 또는 답이 틀렸을 때 각각 RT가 어떻게 되는지 등을 기록하도록 프로그램하여 비교해 볼 수도 있다. 또한 제대로 읽어보지도 않고 답을 적어 넣은 경우를 찾아낼 수도 있으며[15], 부정행위를 한 경우도 찾아낼 수 있다[16].

본 연구에서는 시험자의 총RT와 성적 백분율을 가지고 인공지능기술을 이용하여 분류를 시도하였고, 분류된 각 그룹은 다음과 같은 특징을 가지고 있다. 1그룹 학생은 빨리 정답을 선택하여 학습력과 집중력이 훌륭한 것으로 추측하였다. 다만, 시험문제에 과거 기출 문제가 어느 정도로 출제되었는지 고려하여야 할 것으로 판단된다. 과거 기출문제는 매우 쉬움 또는 쉬움 문항들이 될 수 있으므로 짧은 RT에 영향을 줄 수 있을 것으로 추측된다. 이런 기출 문제의 영향을 고려하더라도 높은 성적과 짧은 RT로 볼 때 1그룹으로 분류된 학생들은 수업을 충분히 이해하고 있으므로 좀 더 높은 단계의 학습을 진행할 수 있도록 유도하면 학생 역량 발전에 도움이 될 것이다. 2그룹 학생은 충분한 시간 동안 문제를 읽고 정답을 선택하여서, 비교적 바람직한 시험행동으로 판단하였다. 신중한 태도로 문제를 정독하고 답을 선택하는 성향인 것으로 추측되므로 현재와 같은 진중한 태도를 유지하면서 조금 빠른 결정을 할 수 있도록 유도하면 좋을 것으로 판단하였다. 3그룹 학생은 문제 풀이 시간은 길었으나 성적은 중하위권이므로, 학습 능력과 방법에 문제가 있는지를 중심으로 학생 상담을 해야 할 것으로 판단하였다. 쉬움 또는 매우 쉬움 문제가 82.6%이었던 것을 고려한다면 충분한 이해를 하지 못하였거나 복습 등의 학습을 하지 않았거나 기출문제를 확인하지 않았을 가능성이 높다. 따라서 평소에 충분히 복습하는 학습 태도를 권하는 것이 좋을 것으로 판단하였다. 4그룹 학생은 시험 문항이 요구하는 지식과 학습목표에 대한 낮은 이해도와 포기하듯 빠른 답 선택(짧은 총RT)을 하는 성향일 것으로 추측된다. 이러한 학생들에게는 부족한 학습 목표 달성을 위한 보충 학습 권유와 더불어 포기하지 않고 정답 선택에 신중한 시험 수행 태도를 가지도록 학업 상담을 해야 할 것으로 판단하였다. 그러나, 상기 4개의 그룹을 나누는 경계선에 있는 학생들은 특정 그룹에 명확히 포함되지 않는 경향이 있으므로 이런 학생들에 대해 상담할 때 이를 충분히 고려해야 할 것으로 판단된다.

현재까지 CBT를 개발하여 실시해 온 대학들은, CBT를 이용하여 시험자가 해당 문항에 소비하는 시간, 답안 수정 여부, 문항 간의 이동 패턴 등 모든 행동 패턴을 실시간으로 기록하여 분석한다면 수험생을 효과적으로 감독 및 지도할 수 있겠다는 초기적 생각을 피력한 바 있다[5]. 그러나 저자들이 아는 한 지금까지 국내에는 이런 항목들을 활용한 실제의 결과는 보고되어 있지 않다. 저

자들은 CBT가 지필시험으로는 불가능한 새로운 지표를 생산해 낼 수 있다는 것을 입증하였으며, 이를 이용하여 학생들의 시험행동을 4개의 합리적인 설명이 가능한 그룹으로 분류할 수 있었다. 이 결과는 맞춤형 성적상담에 활용될 수 있을 것이다.

본 연구의 제한점은 다음과 같다. 우선, 일개 의과대학에서 고부담 과목인 소아과학의 총괄 평가 1회를 대상으로 한 연구의 결과라는 점이다. 한 개의 과목에서 나온 결과를 다른 여러 교과목에 일률적으로 적용하기에는 한계가 있다. 또한 연구 대상이 40명으로 적은 인원수이고, 단 1회의 시험 결과를 이용하였으므로 그 결과로 전체를 대표하기는 충분하지 않다. 따라서 이 연구의 결과가 전체를 설명할 수 있는지에 대하여 추가 연구가 필요하다. 이같이 RT와 성적을 이용하여 분류한 그룹이 실제 상담에서는 어떻게 나타날지에 대한 추가적인 연구도 필요하다.

추후 본 논문과 같은 연구를 진행하고자 할 때는 CBT를 시행하기 전에 미리 시스템에 RT와 같이 추가로 측정하고자 하는 항목을 추출할 수 있는 프로그램을 설치해 두어야 한다. 만약, 충분한 크기의 데이터베이스를 운영할 수 있다면, 다양한 조건을 부여한 RT를 측정하여 학생들의 시험행동을 좀 더 세분하여 분류할 수 있을 것이다.

5. 결론 및 제언

CBT는 평가 문항에 풍부한 시청각 자료를 넣을 수 있고 빠르게 결과를 얻을 수 있는 장점이 있다. 또한 CBT에서는 지필시험에서는 생산할 수 없는 새로운 지표를 추출할 수도 있다. 이 연구에서는 CBT에서 RT를 측정하였고, RT가 시험의 난이도와 비례한다는 것을 보여주었다. 그리고 RT와 성적 백분율을 결합하여 이를 인공지능으로 분석함으로써, 학생들의 시험행동을 합리적인 설명이 가능한 4가지 그룹으로 분류할 수 있었다. RT와 성적 결과를 이용하여 학생들의 시험행동과 성향을 분류하면 학생 개개인에게 맞춤 상담을 제공할 수 있으리라는 가능성을 보여주었다. CBT에서 이 연구에서와 같이 학생의 개인 학습 성향을 측정할 수 있는 인자들을 발굴한다면, 의학교육 성과 평가는 물론 학생 역량 강화에 도움이 될 수 있는 전혀 새로운 측정 도구를 가지게 될 것이다. CBT를 이용한 이러한 결과를 바탕으로 추가적인 후속 연구를 통해 학생 개개인에게 맞춤 상담을 제공할 수도 있으며, 나아가 지필시험으로는 도출하기 어려운 새로운 학습 평가 및 상담 지표를 찾아낼 수 있을 것으로 기대한다.

References

[1] S. Huh, "Application of computerized adaptive testing in medical education", *Korean Journal of Medical Education*, Vol.21, No.2, pp.97-102, Jun. 2009.
DOI: https://doi.org/10.3946/kjme.2009.21.2.97

[2] K. U. Lee, B. K. Kim, S. H. Baik, "A New Approaches of development Systems of Test Items for National Medical Licensing Examinations", *Korean Journal of Medical Education*, Vol.5, No.2, pp.1-10, Apr. 1994.
DOI: https://doi.org/10.3946/kjme.1994.5.2.1

[3] M. K. Yim, The study for development of computer-based national health licensing examination, Korea health personnel licensing examination institute, Korea, RE01-1545-00, pp.92-94.

[4] C. K. Lee, J. S. Park, E. I. Lee, S. J. Lee, E. S. Park, Y. J. Park, "A Comparative Study of Item Analysis by Item Response Theory Based for Initiating CAT(Computer Adaptive Test) System", *Korean Journal of Medical Education*, Vol.13, No.1, pp.107-115, May. 2001.
DOI: https://doi.org/10.3946/kjme.2001.13.1.107

[5] E. J. Im, W. K. Lee, Y. C. Lee, B. H. Choe, S. K. Chung, et. al, "Development of Computer-Based Test (CBT) and Student Recognition Survey on CBT", *Korean Journal of Medical Education*, Vol.20, No.2, pp.145-154, Jun. 2008.
DOI: https://doi.org/10.3946/kjme.2008.20.2.145

[6] J. W. Park, E. C. Jang, J. W. Choi, S. J. Lee, "The Experience of Web-Based Test in Medical Education", *Korean Journal of Medical Education*, Vol.18, No.2, pp.183-192, Aug. 2006.
DOI: https://doi.org/10.3946/kjme.2006.18.2.183

[7] P. C. Kyllonen, J. Zu, "Use of Response Time for Measuring Cognitive Ability", *Journal of Intelligence,* Vol.4, No.4, pp:14. Nov. 2016.
DOI: https://doi.org/10.3390/jintelligence4040014

[8] S. L. Wise, "An Investigation of the Differential Effort Received by Items on a Low-Stakes Computer-Based Test", *Applied Measurement in Education*, Vol.19, No.2, pp.95-114, Jun. 2006.
DOI: http://dx.doi.org/10.1207/s15324818ame1902_2

[9] H-H. Bock, "Clustering Methods: A History of k-Means Algorithms", In: P. Brito, G. Cucumel, P. Bertrand, F. Carvalho, (eds) Selected Contributions in Data Analysis and Classification. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg, 2007, pp.161-172.
DOI: https://doi.org/10.1007/978-3-540-73560-1_15

[10] Y. M. Chae, S. G. Park, I. Park, "The relationship between classical item characteristics and item response time on computer-based testing", *Korean Journal of Medical Education*, Vol.31, No.1, pp.1-9, Mar. 2019.
DOI: http://dx.doi.org/10.3946/kjme.2019.113

[11] S. J. Ko, "Predicting Learning Achievement Using Big

Data Cluster Analysis - Focusing on Longitudinal Study", *Journal of Digital Contents Society*, Vol.19, No.9, pp.1769-1778, Sep. 2018.
DOI: http://doi.org/10.9728/dcs.2018.19.9.1769

[12] K. J. Lee, M. Lee, W. Kim, "Blog Classicification Using K-means", Proceedings of the 11th International Conference on Enterprise Information Systems - Software Agents and Internet Computing, ICEIS 2009, Milan, Italy, pp.61-67, 2009.
DOI: https://doi.org/10.5220/0001949600610067

[13] D. Arthur, S. Vassilvitskii, "K-Means++: The Advantages of Careful Seeding", Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, Society for Industrial and Applied Mathematics, Philadelphia, pp.1027-1035, 2007.

[14] Y. Kang, G. Lee, " A comparison of item selection methods using response times in computerized adaptive testing", *Journal of Educational Evaluation*, Vol.35, No.2, pp.273-298, Jun. 2022.
DOI: https://doi.org/10.31158/JEEV.2022.35.2.273

[15] H. Guo, J. A. Rios, S. Haberman, O. L. Liu, J. Wang, In Paek, "A New Procedure for Detection of Students' Rapid Guessing Responses Using Response Time", *Applied Measurement in Education*, Vol.29, No.3, pp.113-183, Sep. 2016.
DOI: https://doi.org/10.1080/08957347.2016.1171766

[16] S. Sinharay, "Detection of Item Preknowledge Using Response Times", *Applied Psychological Measurement*, Vol.44, No.5, pp.376-392, Jul. 2020.
DOI: https://doi.org/10.1177/0146621620909893

이 미 정(Mee Jeong Lee) [정회원]

- 1996년 2월 : 중앙대학교 의과대학 (의학 학사)
- 2004년 2월 : 울산대학교 대학원 (의학석사)
- 2011년 2월 : 울산대학교 대학원 (의학박사)
- 2004년 3월 ~ 현재 : 단국대학교 의과대학 소아청소년과학교실 교수

〈관심분야〉
의학교육, 소아혈액종양, 아동학대, 사춘기교육

채 유 미(Yoo Mi Chae) [정회원]

- 2001년 2월 : 이화여자대학교 의과대학 (예방의학석사)
- 2009년 2월 : 이화여자대학교 의과대학 (예방의학박사)
- 2007년 3월 ~ 2016년 2월 : 단국대학교병원 직업환경의학과
- 2016년 3월 ~ 현재 : 단국대학교 의과대학 의학교육학교실 교수

〈관심분야〉
교육(과정)평가, 자기주도학습, 평생학습

박 석 건(Seokgun Park) [정회원]

- 1981년 2월 : 서울대학교 의과대학 (의학과 학사)
- 1982년 2월 ~ 1993년 2월 : 포항기독병원, 동국대학교 의과대학, 미국국립보건원
- 1994년 2월 ~ 2020년 2월 : 단국대학교 의과대학 핵의학과 교수
- 2020년 3월 ~ 현재 : 단국대학교 명예교수

〈관심분야〉
의학교육, 의료윤리, 핵의학

박 일 용(Ilyong Park) [정회원]

- 1998년 2월 : 경북대학교 전자전기컴퓨터학부 (전자공학 학사)
- 2000년 2월 : 경북대학교 전자공학과 (전자공학 석사)
- 2004년 8월 : 경북대학교 전자공학과 (전자공학 박사)
- 2008년 3월 ~ 현재 : 단국대학교 의과대학 의공학교실 교수

〈관심분야〉
의용 생체 신호, 의료기기 시스템, 인공지능, 스마트교육