한국심리학회지: 일반
Korean Journal of Psychology: General
2021, Vol. 40, No. 4, 539-566
http://dx.doi.org/10.22257/kjp.2021.12.40.4.539

컴퓨터 기반 적응적 심리 검사 제작을 위한 문항 선정 알고리즘으로서 Alternating Model Tree의 활용 가능성 탐색*

윤 정 한 이 태 헌†

중앙대학교 심리학과

컴퓨터 기반 적응적 검사는 이전에 제시된 문항에 대한 반응을 기반으로 피검사자의 잠재 특질 수준을 추정하기에 가장 적절한 다음 문항을 선택해 출제함으로써 피검사자별 맞춤형 검사를 제시하는 컴퓨터 기반 검사형태다. 컴퓨터 기반 적응적 검사 제작의 핵심 요소 중 하나는 문항 선정 알고리즘이라 할 수 있으며, 최근 결정-트리를 이용한 적응적 심리 검사 구성에 대한 관심과 적용 사례가 늘어나고 있다. 결정-트리는 기계학습 분야에서 주로 연구되어 온 예측 모형 중 하나로서 쉽게 해석 가능한 트리-구조를 가진다는 장점에도 불구하고 과적합 문제에 매우 취약하다는 것이 알려져 있다. 본 연구의 목적은 기계학습 분야에서 결정-트리의 대안으로 제시된 앙상블 모형 중에서 해석 가능한 트리-구조를 지닌 Alternating Model Tree (AMT)가 컴퓨터 기반 적응적 심리 검사 제작에 활용될 수 있는지 탐색하는 데 있다. 이를 위해 먼저 AMT의 작동 방식을 적응적 검사의 특징에 비추어 상술하였고, 검사 점수를 예측하는 AMT 기반의 적응적 검사와 결정-트리 기반의 적응적 검사의 예측 성능을 두 개의 심리 검사를 대상으로 비교하였다. 그 결과, AMT는 적응적 검사의 특징을 가지는 것으로 확인되었고, AMT 기반 적응적 심리 검사의 성능은 결정-트리 기반 적응적 심리 검사의 성능과 유사하거나 더 나은 결과를 보였다. 이러한 결과를 바탕으로 본 연구의 의의와 한계, 후속연구에 대한 제언 등을 논의하였다.

주요어 : 적응적 심리 검사, 컴퓨터 기반 검사, 결정-트리, Alternating Model Tree, 문항 선정 알고리즘

---

* 이 논문은 2019년도 중앙대학교 CAU GRS 지원에 의하여 작성되었음.

† 교신저자: 이태헌, 중앙대학교 사회과학대학 심리학과 부교수, (155-756) 서울시 동작구 흑석로 84
Tel: 02-820-5124, E-mail: lee0267@cau.ac.kr

컴퓨터 기반 적응적 검사(computerized adaptive test; CAT)는 피검사자의 특성 수준(trait level)을 추정하기 위한 유용한 정보를 제공하는 문항들을 적응적(adaptive)으로 제시하는 컴퓨터 기반 검사 형태(Rezaie & Golshan, 2015; Kimura, 2017)로 현재 의학(예, Hsueh et al., 2010), 교육학(예, Istiyono et al., 2020) 및 심리학(예, Hu et al., 2020) 등의 다양한 분야에서 활용되고 있다. 여기서 '적응적'이란, 이전 문항들(previous questions)에 대한 피검사자의 응답에 근거해 해당 피검사자의 특성수준을 추정하기에 가장 유용한 정보를 제공하는(informative) 문항(next question)이 선택되어 제시된다(Bjorner et al., 2007; Winarno & Si, 2018)는 의미이다. 일반적인 컴퓨터 기반 적응적 검사의 검사 과정은 그림 1(Magis et al., 2017)과 같으며, 결과적으로 각 피검사자들은 자신에게 맞춰 제작된 검사를 개별적으로 제시받게 된다.

컴퓨터 기반 적응적 검사의 목적은 검사시행의 효율성과 검사 결과의 정확도를 동시에 확보하는 것이다(Magis et al., 2017). 이를 위해 그림 1과 같은 순서에 따라 각 피검사자 별로 맞춤형 검사(tailored test)를 제공한다. 이러한 컴퓨터 기반 적응적 검사는 지필 검사(paper and pencil test)보다 50%이상 축소된 문항 수로 높은 정확도의 검사 결과를 보인다고 알려져 있다(Gibbons et al., 2016; Carlo et al., 2021). 이러한 장점에 기반해 컴퓨터 기반 적응적 검사는 단축형 검사 개발 방법으로도 활용되고 있으며, 최근 심리검사에 대한 적용이 늘고 있다(Seo et al., 2019; Walter et al., 2007; Hu et al., 2020).

컴퓨터 기반 적응적 검사를 개발하는 가장 일반적인 방법은 문항반응이론모형(item response theory models)에 기반을 둔다(Magis et al., 2017). 문항반응이론 모형은 난이도와 변별도 등의 문항 속성(item characteristics) 및 피검사자의 잠재특질(latent trait) 수준을 추정하기 위해 활용되고 있는 심리측정 모형이다. 문항반응이론 모형은 문항 속성 분석 결과를 바탕으로 문제 은행을 구축하고, 피검사자의 반응 패턴에 따라 해당 피검사자의 잠재특질 수준을 추정하기 위해 가장 유용한 정보를 제공하는 다음 문항을 선정하여 제시하는 '적응적'검사의 기본적인 틀(frameowk)을 제공하였다. 문항반응이론모형에 근거한 컴퓨터 기반 적응적 검사를 IRT-based CAT이라고 부르며, 이는 현재까지 컴퓨터 기반 적응적 검사 구성 방법의 표준으로 활용되고 있다(Babcock & Weiss, 2009; Rezaie & Golshan, 2015).

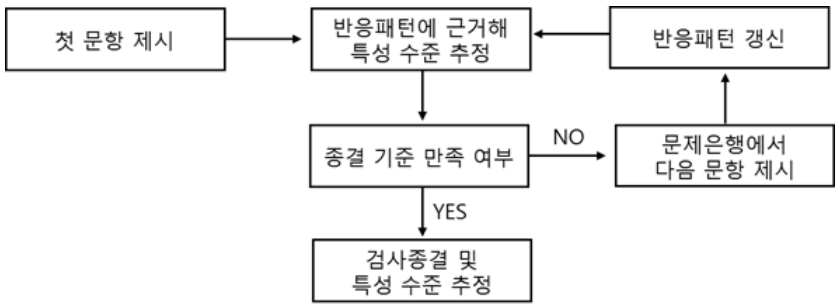

그림 1. 컴퓨터 기반 적응적 검사의 알고리즘

IRT-based CAT은 적응적 검사 제작에서 차지하는 이러한 지위에도 불구하고, 이론적 그리고 현실적으로 극복해야할 문제점을 안고 있다. 먼저, 문제 은행 구축 과정에서 많은 시간과 비용이 요구된다. IRT-based CAT은 모델-기반 접근으로서 문항과 잠재특질 간 관계에서 (단일)차원성(unidimensionality), 지역독립성(local independence), 단조성(monotonicity) 등이 성립한다는 가정을 바탕으로 문항 속성을 분석하고 문제은행을 구축한다. 이 과정에서 사전검사(pretest)를 실시하여 후보 문항들이 문항반응이론 모형에서 요구되는 가정을 충족하는지 확인하는 과정을 반드시 거치게 된다(Antal., 2013). 그러나 실제 데이터에서는 위 가정이 충족되지 않는 경우가 많고, 따라서 다수의 문항이 문제 은행 후보에서 탈락되거나 후보 문항 자체를 재구성해야할 경우가 발생하기도 한다. 이는 문제 은행 구축 과정에서 시간과 비용 그리고 통계적 가정의 성립 여부를 판단할 수 있는 전문적 지식과 경험이 요구된다는 것을 의미한다(Ueno & Songmuang, 2010; Michel et al., 2018; Yan et al., 2004).

IRT-based CAT의 문항 선정 방식 역시 구현을 위해 상당한 전문성과 시간, 비용을 요구한다. IRT-based CAT에서의 문항 선정과정은, 피검사자가 주어진 문항에 응답한 순간 실시간으로 해당 피검사자의 특질 수준 추정치를 계산한 후, 그 정확성을 높이는 데 가장 유용한(informative) 문항을 문제 은행에서 찾아 다음 문항으로 제시하는 하는 과정으로 이루어진다. 이 때 가장 '유용한' 문항을 선정하기 위해 다양한 방법이 제안되어 왔지만 (van der Linden et al., 2009; Thissen & Wainer, 2001; van der Linden et al., 2000; Wainer et al., 2000), 이들은 공통적으로 특질 수준을 추정하는 과정에서 수치적 적분(numerical integration) 계산을 동반하고 있어 구현에 필요한 전문성이 요구되며 또한 실시간 구현을 위한 비용이 발생할 수밖에 없다.

최근에는 컴퓨터 기반 적응적 검사(이하 CAT)을 구현하기 위한 대안적 방법론으로 검사 총점을 예측하는 결정-트리(decision tree; DT)가 활용되고 있다(Michel et al., 2018; Delgado-Gomez et al., 2016; Gibbons et al., 2016; Gibbons et al., 2013; Ueno & Songmuang, 2010; Yan et al., 2004). 결정-트리 기반 적응적 검사(이하, DT-based CAT)는 비모수적 통계 학습(nonparametric statistical learning) 기법의 일종인 결정-트리(Kuhn & Johnson, 2013)를 기반으로 하며, IRT-based CAT과는 달리 (단일)차원성, 지역독립성, 단조성 등의 가정을 요구하지 않는다. 따라서, 가정 성립 여부를 확인하기 위한 문항 분석이 필요하지 않고, 문제은행을 구축하기 위한 절차 역시 요구되지 않는다(Michel et al., 2018). 특히 문항반응이론 모형의 가정이 위배되었을 경우 DT-based CAT이 IRT-based CAT보다 더 높은 검사정확도를 보인다는 점은 주목할 만하다(Yan et al., 2004; Ueno & Songmuang, 2010). 관련하여 DT-based CAT은 검사 총점 예측에 투입되는 문항을 특정 심리 척도에 포함된 문항에 국한할 필요가 없다는 장점을 지닌다. 예를 들면, 피검사자의 과거 이력 등에 대한 정보가 예측에 활용될 수 있으며 이는 검사 총점 예측 정확도 향상에 긍정적인 영향을 미칠 수 있다(Delgado-Gomez et al., 2016). 무엇보다 DT-based CAT의 가장 큰 장점은, 문항 선정 및 피검사자의 특질 수준 추정이 검사 진행 과정 중 실시간으로 이루어질 필요가 없고, 검사 시행 전 완결된 적응적 검사를 트리 구

조를 바탕으로 미리 구성해 놓을 수 있다는 점에 있다(Delgado-Gomez et al., 2019). 이는 적응적 검사의 제작자가 최종 검사의 구조를 미리 검토할 수 있다는 장점도 제공한다(Yan et al., 2016). 트리 구조에 따라 결정된 문항 제시 순서를 CAT으로 구현하는 것 역시 크게 복잡하지 않다. 이러한 장점에 근거해 DT-based CAT을 심리검사에 대해 적용하는 사례도 지속적으로 늘고 있다(예, Gibbons et al., 2013; Delgado-Gomez et al., 2016; Michel et al., 2018).

예측 모형으로서 결정-트리의 성능은 주로 기계 학습(machine learning) 분야에서 연구되었으며, 결정-트리는 과적합(overfitting) 문제, 즉, 주어진 학습데이터[1]에서 보다 새로운 데이터에서 예측 성능이 떨어지는 문제에 매우 취약한 것이 알려져 있다. 이에 기계 학습 분야에서는 결정-트리보다 과적합(overfitting) 문제에서 보다 자유롭고 더 향상된 예측 성능을 보이는 배깅(bagging), 랜덤 포레스트(random forests), 부스팅(boosting) 등의 앙상블 모형(ensemble models)이 개발되어 다양한 분야에서 예측 모델링에 활용되고 있다(James et al., 2013). 이러한 다양한 종류의 앙상블 모형은 기본적으로 수백 혹은 수천 개의 결정-트리로부터 얻을 수 있는 예측값을 (가중) 평균함으로써 최종 예측값을 결정하는 공통점을 가진다. 이러한 앙상블 모형은 결정-트리에 비해 훨씬 더 우수한 예측 성능을 가진 것이 잘 알려져 있음에도 불구하고 CAT 분야에 직접 적용되기는 불가능하다. 왜냐하면 앙상블 모형은 단순하고 해석 가능한 트리-구조를 가지지 못하기 때문이다(Michel et al., 2018; Gibbons et al., 2013; Hastie et al., 2009).

DT-based CAT은 기본적으로 결정-트리를 기반으로 하기 때문에, DT-based CAT 역시 과적합 문제로부터 자유로울 수 없다. 그러나, 앞서 논의하였듯이 DT-based CAT의 장점 또한 명백하다. 따라서, 본 논문의 목적은 DT-based CAT의 장점을 살리되 보다 향상된 예측 성능을 보이는 기계 학습 기반 CAT을 찾는 데 있다. 본 논문에서는 기계학습 분야에서 결정-트리의 대안으로 제시된 앙상블 모형 중에서 해석 가능한 트리-구조를 지닌 Alternating Model Tree(이하 AMT)가 컴퓨터 기반 적응적 검사 제작에 활용될 수 있는지 탐색하고자 한다. 이를 위해 먼저, 결정-트리 모형과 함께 AMT의 구성과정 및 예측값이 산출되는 작동방식을 소개하였다. 다음으로, 실제 수집된 심리검사 자료를 이용하여 AMT를 구성하고 이를 적응적 검사모형으로서 활용할 때 어떠한 과정으로 검사가 진행되는지 예시를 통해 소개하였다. 이를 통해 검사모형으로서의 AMT의 특징과 컴퓨터 기반 적응적 검사를 정의하는 특징들을 비교함으로써 AMT-based CAT이 가능할 수 있음을 보이고자 하였다. 또한, 동일한 학습 데이터를 대상으로 AMT-based CAT과 DT-based CAT의 검사 총점 예측 성능을 비교함으로써 AMT-based CAT의 활용 가능성을 뒷받침하고자 하였다.

1) 학습데이터(training data)는 훈련데이터라고도 부르며, 심리학 연구에서는 모형을 추정하고 적합시키기 위해 연구자가 수집한 표본 데이터(sample data)에 해당한다고 할 수 있다. 현실적으로 표본 데이터 수집 이후 새로운 데이터를 추가로 얻기 어려운 경우가 많기 때문에, 표본 데이터를 나누어 일부는 모형을 추정하는 데 사용하고 나머지를 새로운 데이터로 간주하기도 한다.

가상 데이터 소개

본 논문에서는 결정-트리 모형과 AMT의 구성과정과 피검사자의 결과변수 예측값이 산출되는 모형작동방식을 설명하기 위해 예시 데이터로 가상데이터를 생성했다. 이는 구성개념(construct)에서 특정값을 갖는 피검사자가 리커트 문항에 반응한 것을 모사한 데이터다. 피검사자는 평균이 0이고, 표준편차가 2인 정규분포(normal distribution)로부터 무작위 추출된 구성개념 값을 배정 받았으며 이와 같은 피검사자를 총 5,000명 생성하였다. 그리고 구성개념을 측정하는 문항을 생성하였다. 이때 난이도 및 변별도 등의 문항 특성은 Modified Graded Response Model(Muraki, 1990, 이하 MGRM)에 의해 정의 및 생성되었다. MGRM은 문항반응이론 모형 중 하나로, 심리검사에서 자주 활용되는 리커트 척도와 같이 모든 문항에 동일한 범주형 척도가 제시되는 검사를 분석하기 적절한 모형이다(Magis et al., 2017). 본 논문에서는 통계프로그램 R의 catR 패키지에 존재하는 함수 'genPolymatrix'를 통해 무작위로 총 60개 문항을 만들었다. 각 문항은 MGRM의 파라미터 중 하나인 변별도 모수(discrimination parameter)는 로그-정규 분포(log-normal distribution)로부터, 또 다른 파라미터인 위치모수(location parameter)는 표준정규분포(standard normal distribtution)로부터 무작위로 배정 받았으며, 모든 문항은 5점 리커트 척도로 측정되도록 문항 생성과정을 거쳤다. 이때 변별도 모수의 기술통계치는 표 1, 위치모수의 기술통계치는 표 2에 제기되어 있다. 그리고 이 다음 과정으로 문항 반응 데이터를 생성하였으며, 이는 catR 패키지의 'genPattern'함수를 통해 만들어졌다. 또한 각 피검사자의 검사총점 변수를 따로 만드는 전처리 과정을 거쳤으며, 이후 예시로써 설명하는 결정-트리 모형과 AMT는 모두 이 검사총점을 예측하도록 구성되었다. 총 60개의 가상문항이 문제은행을 이루어 사용가능한 예측변수로 주어졌으며, 검사총점이 결과변수로 지정되었다.

생성된 가상 데이터의 검사총점 기술통계치

표 1. 가상 문항의 변별도 모수 기술통계치

최솟값	최댓값	평균	표준편차	중앙값
0.762	1.275	1.0187	0.105	1.016

표 2. 가상 문항의 위치모수 기술통계치

최솟값	최댓값	평균	표준편차	중앙값
-1.805	2.402	0.118	0.914	0.035

표 3. 가상데이터의 검사총점 기술통계치

최솟값	최댓값	평균	표준편차	중앙값
0.0	240.0	118.5	68.9	117.0

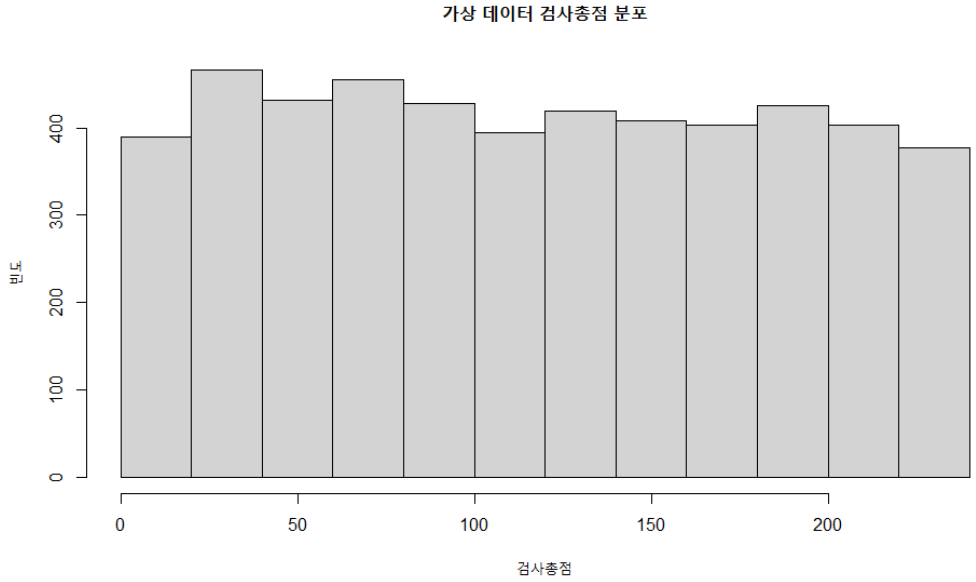

그림 2. 가상데이터의 검사총점 분포

는 표 3에 제시되어 있으며 검사총점의 분포는 그림 2와 같다.

결정-트리

결정-트리(Decision-Tree; DT)는 주어진 학습 데이터를 여러 개의 서로 다른 하위집단으로 분할하는 기계-학습 방법으로, 관측 가능한 결과 변수를 예측하는 것을 목적으로 한다(Song & Ying, 2015).

결정-트리는 일반적으로 모든 학습데이터 피검사자로 이루어진 하나의 집단으로 출발해(Šerbec et al., 2011), 하향식 반복 분할(top-down recursive splitting)을 진행함으로써 구성된다(Myles et al., 2004).

이러한 결정-트리 모형을 구성할 수 있는 대표적인 알고리즘이 Breiman 등(1984)에 의해 제안된 Classification And Regression Trees (CART)로, 이를 통해 분류문제 또는 회귀문제를 다루는 결정-트리를 구성할 수 있다(Myles et al., 2004; Michel et al., 2018). 본 논문에서는 앞서 생성한 가상데이터에 대한 CART의 회귀 결정-트리를 예시로써 결정-트리의 구성과정 및 모형작동방식을 소개하였다.

예시: 회귀 결정-트리 구성과정[2)]

회귀 결정-트리의 구성과정은 그림 3과 같이 주어진 학습데이터의 피검사자들이 모두 속하는 하나의 집단으로 시작한다. 이때 학습데이터의 피검사자들이 속하는 집단을 노드(node)라고 부르며, 첫 번째 노드를 따로 루트 노드(root node)라고 부른다.

2) Using Classification and Regression Trees: A practical primer(Ma, 2018)의 Notation을 차용하였음.

그림 3. 가상 데이터의 루트 노드

그림 3과 같이 노드는 노드 내에 존재하는 학습데이터 피검사자의 관측된 결과변수 평균값이 예측값으로 부여된다. 예를 들어, 그림 3의 루트노드는 노드 내의 5,000명의 피검사자의 관측된 결과변수 평균값 118.5가 예측값이 된다.

CART는 이 루트노드를 두 개의 하위노드로 분할함으로써 모형 구성을 시작한다. 이때 분할은 특정 예측변수의 특정 지점을 기준으로 이루어지며, 예시의 경우 특정 문항의 특정 값을 기준으로 분할이 진행된다. 이때 분할기준은 노드 내 분산(within variance)을 통해 선택되는데, 노드 내 분산이란 노드 내 결과변수 값의 흩어짐 정도를 의미한다. CART에서는 노드 내 분산을 식 (1)과 같이 노드 내 피검사자들로부터 계산한 결과변수에서의 편차제곱합(the sum of squared deviation)으로 계산한다.

$$i(\tau) = \sum (y_i - y_{\cdot})^2 \qquad (1)$$

이때, $i(\tau)$는 특정 노드 τ의 노드 내 분산, y_i는 노드 τ에 속하는 i번째 피검사자의 결과변수 값, $y_{\cdot}$은 노드 내 피검사자들의 결과변수 평균값을 의미한다.

결정-트리 모형은 이 노드 내 분산을 이용해 모든 예측변수의 모든 분할가능지점들 중 하나를 선택해 대상 노드에 대한 분할을 진행한다. 모든 예측변수의 모든 분할지점 별로 분할대상노드를 분할해보는데, 이때 분할 대상이 되는 노드를 따로 부모노드(parent node), 분할을 통해 만들어진 두 개의 하위노드를 따로 자식노드(child nodes)라고 한다. CART는 각 분할가능지점들로부터 부모노드의 노드 내 분산에서 자식노드들의 노드 내 분산을 뺀 불순성 감소량(impurity reduction) 식 (2)를 계산한다.

$$\triangle = i(\tau_{parent}) - i(\tau_{l.ch}) - i(\tau_{r.ch}) \qquad (2)$$

이때, $\triangle$는 노드 내 분산 감소량, $i(\tau_{parent})$는 부모노드의 노드 내 분산, $i(\tau_{l.ch})$는 분할을 통해 생성되는 왼쪽 자식노드의 노드 내 분산, $i(\tau_{r.ch})$는 오른쪽 자식노드의 노드 내 분산를 의미한다.

결정-트리는 각 분할가능지점들로부터 계산한 노드 내 분산 감소량 중 가장 큰 감소량을 보인 예측변수의 지점을 분할기준으로 선택해 분할을 진행한다. 예를 들어, 그림 3의 루트노드를 대상으로 첫 번째 분할을 진행한 경우 Q56의 '1'을 기준으로 분할이 진행된 그림 4와 같은 모형이 나타났다. 이는 그림 3의 루트노드를 대상으로 모든 예측변수의 분할가능지점 중 Q56의 '1'이 분할기준이 되었을 경우 노드 내 분산 감소량이 최대가 되었음을 의미한다. CART의 결정-트리 모형 구성과정은 이러한 분할의 하향식 반복을 통해 이루어진다. 예를 들어, 모형구성의 그 다음 과정으로 그

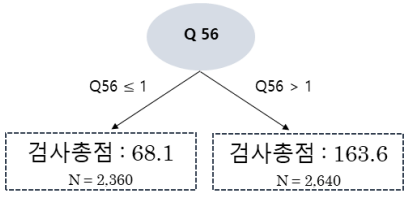

그림 4. 결정-트리 모형의 첫 번째 분할

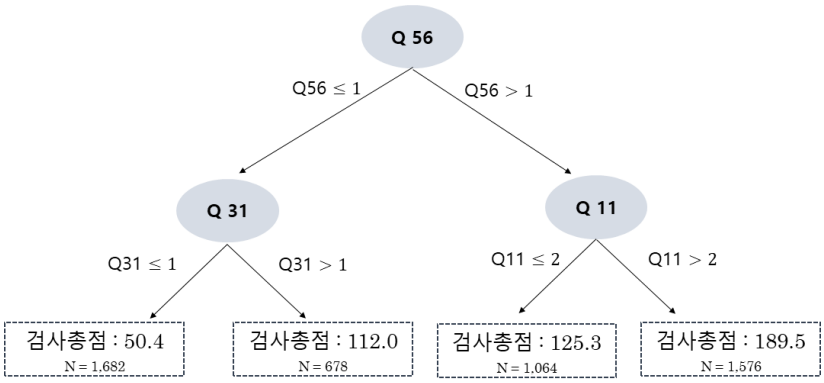

그림 5. 결정-트리 모형의 두 번째 분할

림 4에 나타난 두 개의 자식노드를 부모노드로 두고 이와 같은 분할지점 탐색 및 분할 과정을 반복하여 그림 5와 같은 회귀 결정-트리를 구성할 수 있다.

결정-트리는 이와 같은 하향식 반복 분할 과정을 사전에 지정한 종결규칙(stopping rules)에 도달할 때까지 진행한다(James et al., 2013). 결정-트리 모형을 구성함에 있어 다양한 종결규칙이 사용될 수 있는데, 예를 들어 그림 5는 분할 대상 노드가 되기 위해서 필요한 학습데이터 피검사자 수가 1000명 이상이어야 하며, 모형의 깊이(depth)가 2를 넘을 수 없다는 두 가지 종결규칙을 지정했을 때 분할을 완료한 모형이 된다. 여기서 모형의 깊이는 결정-트리의 크기를 나타내는 하나의 지표로써 루트노드부터 가장 아래의 노드 사이에 존재하는 층의 수를 의미한다. 루트노드의 깊이는 0으로 인식되며, 그러므로 그림 5는 깊이가 2인 결정-트리이다. 추가적으로 모형에서 자식노드를 갖지 않는, 분할되지 않은 노드들을 따로 종결노드(terminal node) 혹은 잎(leaf)이라고 하며, 그림 5는 종결노드가 4개인 결정-트리에 해당한다. 이러한 종결노드의 수 또한 결정-트리의 크기 혹은 복잡도를 나타내는 지표로써 활용될 수 있다.

본 논문에서는 그림 5를 가상데이터에 대한 최종 트리 구조로 두고, 다음 내용으로 결정-트리가 특정 개인에 대해 결과변수 예측값을 산출하는 방식을 설명하고자 한다. 하지만 이는 하나의 예시이며 어떠한 종결 규칙 등을 사용하느냐에 따라 더 크거나 작은 결정-트리를 구성할 수 있다.

회귀 결정-트리의 작동방식

구성을 완료한 결정-트리는 주어진 예측변수에 대한 반응을 통해 결과변수 값을 예측하는데 사용할 수 있다. 결정-트리 모형으로 결과변수 값을 예측하고자 하는 각 피검사자는 모형의 가장 위에 존재하는 루트노드에서부터 모형이 요구하는 예측변수에 대해 반응한다. 이때, 앞 예측변수에 대한 반응에 따라 다음 노드로 이동하게 되며 그렇기 때문에 피검사자들은 서로 다른 예측변수를 배정받을 수 있다. 즉, 결정-트리 모형에서 각 피검사자들은 자신에게 해당되는 하나의 길(path)을 따라 모

형의 아래로 내려가며, 최종적으로 하나의 종결노드에 도착한다. 그리고 결정-트리 모형은 그 종결노드의 예측값을 그 피검사자에 대한 결과변수 추정치로 제시한다. 예를 들어, 그림 5의 결정-트리 모형을 통해 특정 피검사자에 대한 결과변수 값을 예측한다면 먼저 이 피검사자는 Q56에 대해 반응해야 한다. 이때 이에 대한 반응은 다음으로 이동해야할 노드를 결정한다. 즉, Q56에 1이하의 반응을 보일 경우 Q31을 제시하는 노드로 이동하며, 1 초과의 반응을 보일 경우 Q11을 제시하는 노드로 이동한다. 만약 그림 5에서 피검사자가 Q50에 1 이하의 반응을 보이고, 다음으로 Q31을 출제받아 이에 대해 1 초과의 반응을 보인다면 피검사자는 예측값으로 '112.0'을 갖는 그림 5의 종결노드에 도달하게 된다. 그림 5의 결정-트리는 이와 같은 피검사자에게 모형 구성 과정에서 이 종결노드에 부여된 예측값 '112.0'을 결과변수 예측값으로 제시한다.

이러한 결정-트리 모형의 모형작동방식은 컴퓨터 기반 적응적 검사의 원리와 동일하다(Delgado-Gomez et al., 2016). 그러므로 그림 5와 같이 예측변수가 문제은행의 문항들, 결과변수가 검사의 합성점수로 주어진 결정-트리 모형은 컴퓨터 기반 적응적 검사모형으로서 활용될 수 있다. 그리고 앞서 언급한 바와 같이 컴퓨터 기반 적응적 검사모형으로서의 결정-트리 모형은 여러 장점을 가지며, 이미 많은 문헌들(예, van der Oest et al., 2020; Peute et al., 2020)에서 이를 이용한 적응적 검사 개발 연구가 이루어지고 있다.

결정-트리의 과적합 문제

통계적 학습 모델이 주어진 학습 데이터에 존재하는 노이즈까지 학습함으로써 새로운 데이터가 주어졌을 때 예측 정확도가 하락하는 경우 과적합(overfitting)이 발생했다고 하며, 과적합 문제를 보이는 학습 모델의 성능은 일반화가능성(generalizability)이 낮다고 한다(Brownlee, 2016; Maimon & Rokach, 2014).

결정-트리에서의 과적합(overfitting)은 일반적으로 주어진 학습데이터 크기에 비해 너무 많은 분할을 진행함으로써 발생한다(Maimon & Rokach, 2014). 결정-트리는 주어진 학습데이터에 대한 하향식 반복 분할을 통해 구성되며, 분할을 진행하면 할수록 트리-구조의 복잡도가 증가함을 의미한다. 트리의 구조가 더 복잡해질수록 주어진 학습 데이터에만 존재하는 독특한 특성까지 학습할 가능성이 높아지고, 이는 결국 새로운 자료에서의 예측오차 증가로 이어지게 된다(Sug, 2009; Domingos, 2000; Dietterich & Kong, 1995).

과적합의 발생을 막기 위해 결정-트리는 몇몇의 방법을 사용해 트리의 복잡도에 제약을 가한다. 첫 번째 방법은 임의적인 모형 구성 종결 규칙을 사용하는 것으로, 예를 들어 다음 분할 대상이 될 수 있는 집단은 일정 수의 피검사자를 갖고 있어야 된다는 규칙 하에서 모형을 구성할 수 있다. 두 번째는 가지치기(pruning)라는 방법을 통해 최적의 트리 크기를 추정하는 것으로(Wu et al., 2016), 예를 들어 CART에서는 비용-복잡도 측정치(cost-complexity measure)을 이용해 가지치기를 진행한다(Ma, 2018).

그러나 이러한 방법들은 결정-트리의 과적합 문제에 대한 근본적인 해결책을 제공하지는 못하였으며, 결정-트리에 비해 과적합 문제에서 훨씬 더 자유롭고 예측력이 높은 배깅(bagging), 랜덤 포레스트(random forests), 부스

팅(boosting) 등의 앙상블 기법이 개발되면서 현재는 다양한 분야에서 결정-트리 대신 앙상블 모델이 예측 모델링에 활용되고 있다(James et al., 2013; Hastie et al., 2009).

앞에서도 언급하였듯이 이러한 앙상블 모형은 결정-트리에 비해 훨씬 더 우수한 예측 성능을 보이는 것이 잘 알려져 있지만, 단순 명확하게 해석 가능한 트리-구조를 상실함으로써 CAT에 직접 적용하기가 불가능하다. 따라서, 본 논문에서는 기계학습 분야에서 결정-트리 모형의 대안으로 제시된 앙상블 모형 중에서 해석 가능한 트리-구조를 지닌 AMT가 컴퓨터 기반 적응적 검사 제작에 활용될 수 있는지 탐색하고자 한다. 아래에서는 AMT의 구성과정 및 예측값이 산출되는 작동방식을 소개하였다.

Alternating Model Tree[3)]

AMT는 Eibe Frank(2015) 등에 의해 제안된 회귀문제를 다루는 새로운 결정-트리로, 여러 개의 기본모형(base learner)을 결합한 가산모형(additive model)이다. 그러므로 AMT는 식 (3)과 같이 K개의 기본모형 또는 기본모형의 예측값 $f_1, \ldots, f_K$ 가 결합된 가산모형 또는 가산모형의 예측값 F_K 로 표현할 수 있다. 이때 $\overrightarrow{x_i}$는 i번째 피검사자의 예측변수 벡터(vector)를 의미한다.

$$F_K(\overrightarrow{x_i}) = \sum_{j=1}^{K} f_j(\overrightarrow{x_i}) \quad (3)$$

3) 문헌'Alternating Model Trees'(Frank et al., 2015)의 Notation을 차용하였음.

AMT는 예측값과 학습데이터의 결과변수 관측값 사이의 오차제곱(the squared errors)을 최소화하는 것을 목표로 구성되며, 오차제곱은 식 (4)와 같다. 이때 n은 학습데이터에 존재하는 피검사자의 수를, y_i는 i번째 피검사자의 결과변수 관측값을 의미한다. AMT는 앙상블 기법 중 Forward Stagewise Additive Modeling을 통해 구성된다.

$$\sum_{i=1}^{n} (F_K(\overrightarrow{x_i}) - y_i)^2 \quad (4)$$

Forward Stagewise Additive Modeling은 기본모형을 구성해 이를 현재 모형에 추가함으로써 순차적이고 반복적으로 모형의 예측값을 교정하는 가산모형 구성 모델링이다. 이때 기본모형은 현재의 가산모형이 학습데이터에 대해 갖는 잔차집합(set of residuals)을 통해 구성된다. 예를 들어, 가산모형 F_K에 추가할 K+1번째 기본모형 f_{K+1}을 구성한다면 먼저 현재의 가산모형이 학습데이터에 대해 갖는 잔차집합을 식 (5)와 같이 계산한다.

$$Residuals = \{y_1 - F_K, \ldots, y_n - F_K\} \quad (5)$$

그리고 이 잔차집합을 예측하는 기본모형 f_{K+1}을 구성하는데, 이때 잔차와 기본모형의 예측값 사이의 오차제곱이 최소가 되도록 만들어진다. 이와 같은 K+1번째 기본모형이 가산모형에 추가되며, 이러한 모델링은 학습데이터에 대한 식 (6)의 오차제곱을 최소화한다.

$$\sum_{i=1}^{n} ((y_i - F_K(\overrightarrow{x_i})) - f_{K+1}(\overrightarrow{x_i}))^2 = \sum_{i=1}^{n} (y_i - F_{K+1}(\overrightarrow{x_i}))^2 \quad (6)$$

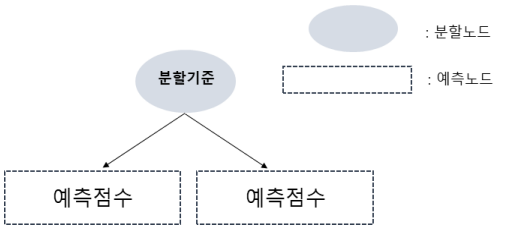

그림 6. Decision Stump

AMT는 이러한 원리 하에 구성되며, 기본모형으로 한 번의 분할만 진행된 결정-트리인 Decision Stump를 사용한다. Decision Stump는 그림 6과 같이 분할 대상 노드의 분할 기준을 가지는 한 개의 분할노드(splitter node)와 분할로부터 나온 자식노드이면서 예측값을 가지는 두 개의 예측노드(prediction nodes)로 이루어져 있다.

앞서 결정-트리를 적용한 가상데이터를 예시로 AMT의 구성과정 및 작동방식을 소개한다.

예시 : AMT 구성과정

먼저, AMT는 그림 7과 같이 학습데이터의 모든 피검사자가 속하는 하나의 예측노드로 구성을 시작한다. 이때, 이 예측노드는 소속되어 있는 피검사자의 결과변수 평균을 예측값으로 가진다. 이를 식 (7)과 같이 첫 번째 기본모형으로 사용해 식 (8)과 같이 학습데이터로 주어진 가상데이터의 피검사자 5,000명에 대한 첫 번째 잔차집합을 계산할 수 있다.

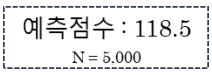

그림 7. AMT의 첫 번째 예측노드

$$F_1(\overrightarrow{x_i}) = 118.5 \tag{7}$$

$$Residuals = \{y_1 - F_1, \ldots, y_{5000} - F_1\} \tag{8}$$

다음으로, 예측노드 중 하나를 대상으로 그 노드의 피검사자에 대한 잔차집합을 예측하는 Decision Stump를 구성한다. 이때, 분할은 계산상의 편의를 위해 특정 예측변수의 중앙값(median value)을 기준으로 이루어진다. 그리고 분할 대상 예측노드와 이에 대한 분할 기준은 학습데이터에 대한 오차제곱이 최소화되도록 선택된다. 즉, 각 예측노드를 대상으로 각 예측변수의 중앙값이 갖는 분할 이후의 오차제곱 감소량(the reduction of squared error)을 모두 계산하고, 그 중 가장 큰 감소량을 보인 예측노드의 분할 기준으로 Decision Stump가 구성된다. 그림 7에 하나의 Decision Stump를 추가하면 그림 8과 같아졌는데, 이는 첫 번째 예측노드를 대상으로 Q56의 중앙값 '3'을 기준으로 분할을 진행했을 때 가장 큰 오차 제곱 감소량이 나타났음을 의미한다. 그리고 이에 따라 두 개의 새로운 예측노드가 생성되었다. 이때 AMT의 Decision Stump가 갖는 각 예측노드에는 잔차집합에 대한 평균값이 아니라 특정 예측변수를 활용한 단순선형회귀모형

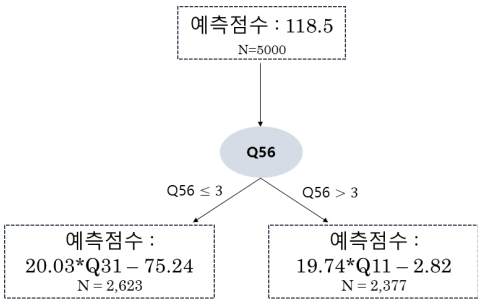

그림 8. 첫 Decision Stump가 추가된 AMT

(simple linear regression model)이 예측값으로 부여된다. 이는 학습데이터에 대한 오차제곱이 최소가 되도록 구성된다.

다음으로, 첫 번째 Decision Stump이자 두 번째 기본모형이 추가됨으로써 AMT가 학습데이터에 대한 예측값이 식 (9)와 같이 달라지며 이에 따라 식 (10)과 같이 잔차집합을 갱신한다.

$$F_2(\overrightarrow{x_i}) = 118.5 + \lambda f_2(\overrightarrow{x_i}) \quad (9)$$

$$Residuals = \{y_1 - F_2, \ldots, y_{5000} - F_2\} \quad (10)$$

이때, 첫 번째 기본모형을 제외한 기본모형, 즉 모든 Decision Stump의 예측값에는 Shrinkage Parameter(λ)라는 하이퍼-파라미터 값이 곱해진다. 이는 과적합 발생을 막기 위한 파라미터로 0부터 1사이의 값을 가질 수 있다. 이 파라미터의 값은 학습데이터에 대한 교차검증 등을 통해 추정할 수 있다. 하지만 본 예시에서는 이 파라미터를 임의로 지정했을 때 AMT의 특징을 보이기에 적절한 결과가 산출되었기 때문에 따로 추정하지 않고 임의의 값 1을 지정하였다.

계속해서, 앞서 갱신된 잔차집합을 예측하는 두 번째 Decision Stump를 구성할 수 있는데, 이때 AMT 내에 존재하는 모든 예측노드 중 하나를 대상으로 구성할 수 있다. 즉, 이미 Decision Stump 구성의 대상으로 사용된 예측노드라도 학습데이터에 대한 오차제곱을 최소화한다면 다시 분할 대상 예측노드로 활용될 수 있다. 예를 들어, 그림 8에 새로운 Decision Stump를 추가하면 그림 9와 같았다. 그림 9를 보면 이미 그림 8에서 분할된 첫 번째 예측노드를 대상으로 새로운 Decision Stump가 구성되었다. 이와 같이 AMT에서는 하나의 예측노드를 대상으로 여러 번의 Decision Stump 구성이 가능하며, 각 예측노드가 분할 가능 대상에서 탈락하지 않기 때문에 Decision Stump가 늘어날수록 다음 분할 대상으로 고려되는 예측노드의 수는 계속 증가한다.

이처럼 AMT는 분할 대상 예측노드 및 분할지점 탐색, Decision Stump 구성, 잔차집합 갱신이라는 일련의 과정을 정해진 수만큼 반복한다. 예를 들어, 그림 9에 세 번째 일련의 과정 진행한다면 그림 10과 같이 3개의 Decision Stump가 결합된 AMT를 구성할 수 있었다. 이때, 그림 10에 추가된 세 번째 Decision Stump가 '20.03*Q31 - 75.24'를 예측값으로 갖는 예측노드를 대상으로 구성되었는

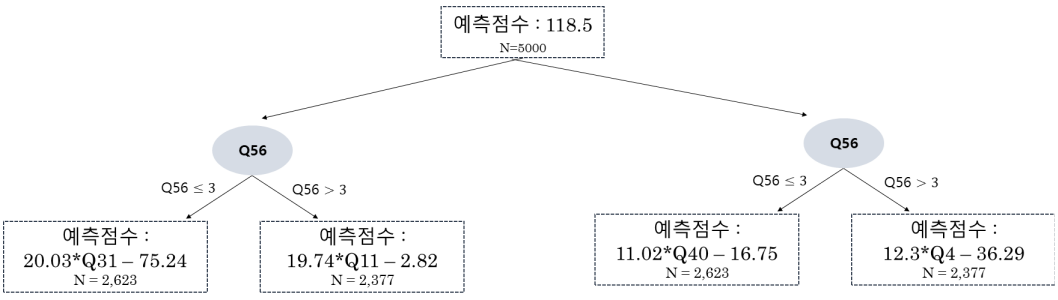

그림 9. 두 번째 Decision Stump가 추가된 AMT

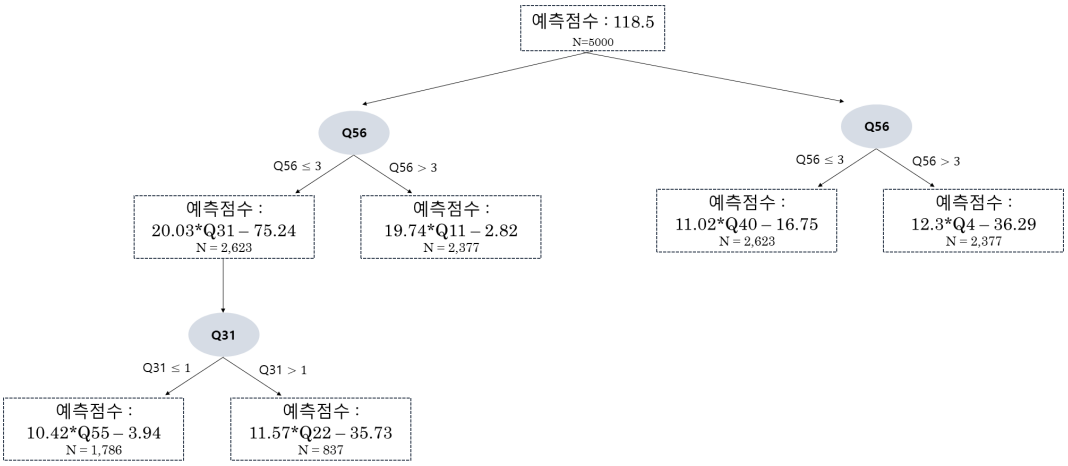

그림 10. 세 번째 Decision Stump가 추가된 AMT

데, 이에 따른 잔차집합의 갱신은 분할 대상이 된 예측노드의 피검사자 2,623명에 대해서만 발생한다.

본 논문에서는 그림 10을 가상데이터에 대한 최종 AMT로 두고, 다음으로 AMT가 특정 개인에 대해 결과변수 예측값을 산출하는 작동방식을 설명하고자 한다.

AMT의 작동방식

AMT를 통해 결과변수를 예측하고자 할 때, 피검사자는 AMT의 가장 상단에 위치한 첫 번째 예측노드에 속하는 것으로 시작한다. 그리고 피검사자는 AMT에서 자신이 속하는 예측노드로부터 만들어진 모든 Decision Stump를 거치며, 이로부터 제시되는 예측변수들에 반응해야 한다.

예를 들어, 그림 10의 AMT를 통해 특정 피검사자의 검사총점을 예측한다면 이 피검사자는 AMT의 가장 상단에 존재하는, 예측점수로 118.5의 값을 갖는 첫 번째 예측노드에 속하는 것으로 시작한다. 그리고 이로부터 파생된 두 개의 Decision Stump를 거친다. 즉, 왼쪽 Decision Stump로 먼저 이동해 처음으로 Q56를 제시받게 된다. 만약 제시된 Q56에 대해 3 이하의 반응을 보인다면 피검사자는 '20.03*Q31 - 75.24'를 예측값으로 갖는 왼쪽 예측노드에 속해지게 되며, 이로부터 Q31이 다음 예측변수로 제시받는다. 계속해서 이 피검사자는 자신이 속한 첫 번째 예측노드로부터 파생된 오른쪽 Decision Stump 또한 거치게 되는데 공교롭게도 이 Decision Stump에서 처음으로 제시 하는 예측변수가 Q56였으며, 앞서 이에 대해 3 이하의 반응을 보였으므로 '11.02*Q40 - 16.75'를 예측값으로 갖는 왼쪽 예측노드에 속하게 된다. 그러므로 피검사자는 Q40을 다음 예측변수로 제시받는다. 그리고 이 피검사자는 앞서 '20.03*Q31 - 75.24'를 예측값으로 갖는 예측노드에 속했기 때문에 이로부터 파생된 Decision Stump 또한 거쳐야 된다. 그러므로 예측변수 Q31를 다시 받게 되며, 만약 이에 대한 반응이 1을 초과했다면 '11.57*Q22

-35.73’을 예측값으로 갖는 오른쪽 예측노드에 속하게 되어 Q22을 다음 예측변수로 제시받게 된다.

이와 같이 특정 피검사자가 AMT에서 자신에게 해당하는 Decision Stump와 예측변수에 대한 반응을 모두 거치게 되면, 피검사자가 속하게 된 모든 예측노드의 예측값을 Shrinkage Parameter를 곱하여 조정한 뒤 합산함으로써 결과변수에 대한 최종 예측값을 계산한다. 이 때, 첫 예측 노드의 예측값은 조정하지 않는다. 예를 들어, 앞선 예시에서의 피검사자에 대한 최종 예측값은 아래 식 (11)과 같이 구할 수 있다.

$$118.5 + \lambda(20.03^{*}Q31 - 75.24) + \lambda(11.02^{*}Q40 - 16.75) + \lambda(11.57^{*}Q22 - 35.73) \quad (11)$$

이처럼 AMT는 속하는 예측노드와 그로부터 파생된 Decision Stump에 따라 피검사자가 여러 길을 가야한다는 점에서 결정-트리와 차이점을 갖는다. 하지만 AMT에서도 앞 예측변수에 대한 반응에 따라 다음에 반응할 예측변수가 서로 다를 수 있다. 또한 자신에게 해당하는 특정 예측변수에 대한 반응만으로 결과변수의 값을 예측할 수 있다.

방 법

본 논문의 목적은 AMT가 컴퓨터 기반 적응적 심리검사 제작 도구로써 활용될 수 있고, AMT-based CAT이 DT-based CAT에 필적하는 혹은 보다 더 뛰어난 검사정확도를 보일 수 있는지 탐색하는 데 있다.

이를 위해 본 연구에서는 하나 이상의 검사총점을 사용하는 두 개의 심리검사를 분석대상으로 선정하였으며, 이에 대한 데이터는 웹사이트 ‘Open-Source Psychometrics Project (https://openpsychometrics.org/)’에서 다운로드하였다.

Narcissistic Personality Inventory

Narcissistic Personality Inventory는 Raskin과 Hall(1979)에 의해 개발된 나르시시즘 특성(narcissism trait) 수준을 측정하는 심리검사로, 이를 단축한 NPI-40(Raskin & Terry, 1988)이 개발되어 사용되고 있다. NPI-40은 총 40개의 문항으로 구성되어 있으며, 각 문항은 두 개의 진술문을 제시한다. 이때 피검사자는 자신의 성격을 가장 잘 반영하는 하나의 진술문을 선택하게 된다. NPI-40은 하나의 검사총점을 사용하며, 이는 0점에서 40점 사이의 값을 가진다. 검사총점이 높을수록 더 높은 수준의 나르시시즘 특성을 의미한다.(Twenge et al., 2008).

본 연구에서는 웹사이트 ‘Open-Source Psychometrics Project’에서 2012년 9월 6일에 업데이트된 데이터를 사용했다. 데이터는 NPI-40의 문항에 대한 응답, 성별, 나이, 반응시간, 검사총점으로 구성되어 있다. 또한 총 11,243명의 피검사자로 이루어져 있었으며, 이 중 (1). 성별을 제대로 입력하지 않은 피검사자, (2). 나이를 14세 미만 또는 80세 초과로 기록한 피검사자, (3). 문항에 대한 반응에서 결측치(missing value)를 보인 피검사자들을 연구에서 탈락시켰다. 903명의 피검사자가 탈락했으며, 총 10,340명의 피검사자로 이루어진 데이터를 이용해 연구를 진행했다. 이때,

NPI-40의 40개 문항으로 구성된 문제은행이 예측변수로, NPI-40의 검사총점이 결과변수로 주어졌다.

Empathizing-Systemizing Test

Empathizing Systemizing Test(이하 EQSQ)는 Empathizing-Sytemizing Theory(이하 E-S theory)에 기반한 심리검사다. E-S theory는 개인의 인지유형(cognitive style)을 두 개의 차원으로 정의한다(Lai et al., 2012). 첫 번째 차원은 공감능력(empathizing)으로, 이는 타인의 감정(emotion)과 생각(thought)을 인식하고 적절한 감정으로 반응하고자 하는 추동(drive)으로 정의된다(Baron-Cohen et al., 2003). 두 번째 차원은 체계화능력(systemizing)으로, 이는 체계(system)의 변수들을 분석하고 체계를 지배하는 규칙들을 끌어내고자 하는 추동으로 정의된다(Baron-Cohen et al., 2003). EQSQ는 공감능력과 체계화능력을 측정하기 위해 설계된 심리검사다(Baron-Cohen et al., 2003; Baron-Cohen & Wheelwright, 2004). 검사는 인지능력을 측정하기 위한 두 개의 하위척도 Empathy Quotient와 Systemizing Quotient로 이루어져 있다. 두 개의 하위척도는 각각 60개 문항으로 이루어져 있으며, 문항들은 리커트 척도로 측정된다. 또한 각 하위척도를 구성하는 60개 문항은 특정 인지능력을 측정하는 문항 40개와 통제 문항 20개로 이루어져 있는데, 통제 문항은 공감능력 또는 체계화능력에 과도하게 몰입하는 것을 방해하기 위해 제시된다(Baron-Cohen et al., 2003). EQSQ는 각 하위척도로부터 검사총점을 하나씩 계산한다. 이때 특정 인지능력을 측정하기 위해 설계된 40개 문항만이 검사총점 계산에 사용된다. 각 척도의 검사총점은 0점부터 80점 사이의 값을 가질 수 있으며, 점수가 높을수록 더 높은 인지능력 발달을 의미한다.

본 연구에서는 웹사이트 'Open-Source Psychometrics Project'에서 2012년 7월 16일에 업데이트 된 데이터를 사용했다. 데이터는 EQSQ의 120개 문항에 대한 응답과 성별, 나이로 구성되어 있다. 그리고 총 13,256명의 피검사자로 이루어져 있는데, 이 중 (1). 성별을 제대로 입력하지 않은 피검사자, (2). 나이를 14세 미만 또는 80세 초과로 기록한 피검사자, (3). 문항에 대한 반응에서 결측치를 보인 피검사자들을 분석대상에서 탈락시켰다. 총 1,801명의 피검사자가 탈락했으며, 최종적으로 11,455명의 피검사자로 이루어진 데이터를 구성할 수 있었다. 이후, 데이터를 각 하위척도별로 나누어 독립적인 분석을 진행했다. 이때, 각 하위척도에서 검사총점 계산에 사용되는 40개 문항을 문제은행으로 사용했으며, 검사총점에 기여하지 않는 통제문항들은 문제은행에서 제외되었다. 그리고 각 하위척도의 검사총점이 결과변수로 주어졌다.

첫 번째 연구방법

첫 번째 연구의 목표는 AMT가 컴퓨터 기반 적응적 검사 제작에 활용될 수 있음을 보이는 것이다. 이를 위해 앞서 소개한 NPI-40을 대상으로 검사모형으로서의 AMT를 구성했다. 총 10,340명의 피검사자들로 구성된 NPI-40 데이터를 학습데이터로 사용해 AMT를 구성했으며, 이때의 Decision Stump의 수를 임의로 8개로 고정하였다. 그리고 또 다른 하이퍼-파라미터인 Shrinkage Parameter는 주어진 학습데이터에서의 10-fold Cross Validation을 통해 추정한 값 0.5로 입력한 경우와 임의의 값

1로 지정한 경우로 나누어 AMT를 구성하였다. 10-fold Cross Validation은 주어진 데이터를 동일한 크기를 가지는 열 개의 부분으로 나누어, 그 중 아홉 개를 합쳐 학습데이터로 사용하고, 나머지 한 부분을 새로운 데이터로 사용하는 교차검증방법이다. 어느 부분을 새로운 데이터로 선택하느냐에 따라 결과가 조금씩 달라질 수 있는데, 일반적으로 서로 다른 부분을 새로운 데이터로 사용하여 얻은 여러 결과의 평균을 사용한다.

두 가지 경우로 나눈 이유는 각 경우가 NPI-40에 대한 서로 다른 형태의 AMT를 보였기 때문이다. 전자의 경우 AMT의 모든 Decision Stump가 첫 번째 예측노드로부터 파생된 형태가 나타난 반면, 후자의 경우 첫 번째 예측노드가 아닌 다른 예측노드에서 파생된 Decision Stump가 포함된 형태가 나타났다. 두 형태 모두 실제로 나타날 수 있는 AMT이며, AMT가 컴퓨터 기반 적응적 심리검사 제작 도구로써 활용될 수 있음을 보이기 위해서는 두 경우 모두 적응적 검사를 정의하는 특징들을 갖고 있음을 확인할 필요가 있었다.

이와 같이 구성된 AMT로 검사를 진행하고자 할 때, 어떠한 과정을 통해 검사가 진행되는지 여러 가상의 예시들을 통해 보이고자 했다. 그리고 이를 통해 검사모형으로서의 AMT가 컴퓨터 기반 적응적 검사를 정의하는 중요한 특징들을 가지며, 이로 인해 AMT-based CAT이 가능함을 보이고자 했다.

두 번째 연구방법

두 번째 연구의 목표는 AMT-based CAT의 검사정확도에 대한 기초 연구 결과를 확보하는 것이다. 이때 검사정확도는 검사의 예측값과 관측된 검사총점 사이의 차이를 의미하며, 연구 목적을 위해 앞서 소개한 두 가지 심리검사를 대상으로 동일한 학습데이터가 주어진 상황에서 DT-based CAT과 AMT-based CAT을 구성하고 검사정확도를 비교했다. 연구는 다음의 단계적 과정을 통해 진행했다.

첫 번째로, 주어진 심리검사 데이터를 적응적 검사모형 구성에 사용할 학습데이터와 이를 평가하는데 사용할 검사데이터(test data)로 무작위 추출을 통해 나누었다. 이때 검사데이터는 모형구성에 사용하지 않은, 모형을 평가하는데 활용할 데이터를 의미한다. NPI-40에서의 학습데이터는 총 5,170명의 피검사자로 이루어졌으며 검사 데이터 또한 동일한 수의 피검사자로 구성되었다. 반면에 EQSQ에서의 학습데이터는 총 5,727명의 피검사자로 이루어졌으며, 검사데이터는 5,728명의 피검사자로 구성되었다.

두 번째로, 동일한 학습데이터를 통해 DT-based CAT과 AMT-based CAT을 구성하였다. 이때 두 검사모형은 동일한 문항 수를 출제하도록 구성되었다. 즉, 동일한 검사 시행 효율성 상황에서 검사정확도를 비교했다는 것이며, 본 논문에서 언급하는 '출제 문항 수'는 검사가 특정 피검사자에게 출제할 수 있는 최대 문항 수를 의미한다. 각 심리척도에서 DT-based CAT과 AMT-based CAT은 출제 문항 수가 2개인 경우부터 문제은행 문항의 과반수인 20개인 경우까지 구성되었다. 이때, 과적합 문제가 발생하는 것을 막기 위한 결정-트리와 AMT의 하이퍼-파라미터들을 주어진 학습데이터에서의 10-fold Cross-Validation을 통해 추정하였다. 이를 통해, 결정-트리의 '분할 대상이 되기 위한 노드의 최소 피검사자 수(minsplit)'와 '복잡도-파라미터(cp)', AMT의 'Shrinkage

Parameter (H)'가 정해졌다.

세 번째로, 위의 과정을 통해 구성된 DT-based CAT과 AMT-based CAT의 검사정확도를 검사데이터를 통해 평가했다. 검사정확도는 검사의 예측값과 피검사자의 관측값 사이에서 계산되는 식 (12)의 상관계수(pearson correlation coefficient), 식 (13)의 MAE(mean absolute error), 식 (14)의 RMSE(root mean squared error)로 나타내었다.

$$Corr = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}} \quad (12)$$

$$MAE = \frac{\sum_{i=1}^{n}|\hat{y}_i-y_i|}{n} \quad (13)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{n}} \quad (14)$$

이때, y_i는 i번째 피검사자의 관찰된 검사총점, $\bar{y}$는 모든 피검사자의 관찰된 검사총점 평균, $\hat{y}_i$는 i번째 피검사자에 대한 적응적 검사모형의 예측값, $\bar{\hat{y}}$는 모든 피검사자의 예측값 평균, n은 총 피검사자 수를 의미한다.

결 과

첫 번째 연구

앞서 언급한 바와 같이 NPI-40 데이터를 대상으로 Shrinkage Parameter를 임의의 값 1로 입력한 경우(그림 11)와 학습데이터에 대한 10-fold Cross Validation을 통해 추정한 값 0.5를 입력한 경우(그림 12)로 나누어 AMT를 구성하였다. 두 경우는 모든 Decision Stump가 첫 번째 예측노드로부터 파생되었는지 여부에서 차이를 보이고 있다. 또한 검사과정 설명의 편의를 위해 각각의 AMT에 존재하는 8개의 Decision Stump에 ①부터 ⑧까지의 번호를 매겼고, 17개의 예측노드에 (a)부터 (q)까지의 번호를 부여했다.

먼저, 그림 11의 AMT를 검사모형으로 활용할 때 모든 피검사자는 예측노드(a)에 속하는 것으로 검사를 시작한다. 그러므로 모든 피검사자는 예측노드(a)로부터 파생된 Decision Stump ①, ②, ③, ④, ⑤, ⑥을 모두 거쳐야한다. 이와 같을 때 발생할 수 있는 가상 피검사자 A와 B의 검사과정은 표 4와 같다.

다음으로, 그림 12의 AMT를 검사모형으로 활용할 때 모든 피검사자는 예측노드(a)로부터 파생된 Decision Stump ①, ②, ③, ④, ⑤, ⑥, ⑦, ⑧을 모두 거쳐야 한다. 이와 같을 때 발생할 수 있는 가상 피검사자 C와 D의 검사과정은 표 5와 같다.

두 가지 형태의 AMT를 통해 검사를 진행할 때 네 명의 가상 피검사자가 거친 검사과정을 소개했다. 이로부터 검사모형으로서의 AMT에 대해 알 수 있는 바는 다음과 같다.

첫 번째로, 검사모형으로서의 AMT는 각 피검사자 별로 맞춤형 검사를 제공한다. AMT에서 모든 피검사자는 첫 예측노드에 반드시 속하며 이로부터 파생된 모든 Decision Stump를 거친다. 예를 들어, 그림 11에서는 첫 예측노드로부터 파생된 6개의 Decision Stump를, 그림 12에서는 첫 예측노드로부터 파생된 8개의 Decision Stump를 모든 피검사자가 거쳤다. 그

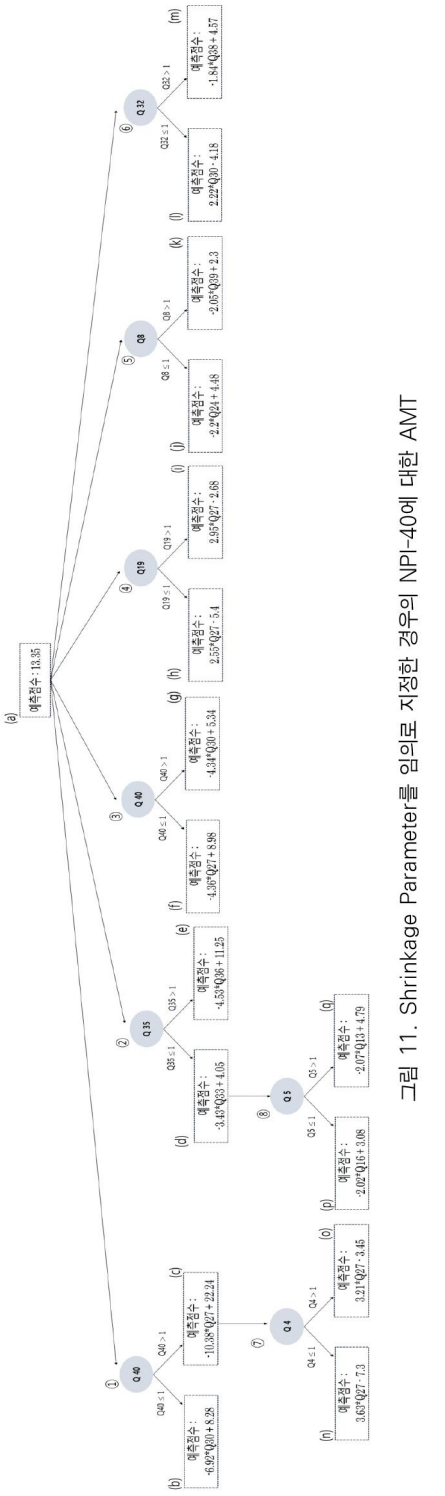

그림 11. Shrinkage Parameter를 임의로 지정한 경우의 NPI-40에 대한 AMT

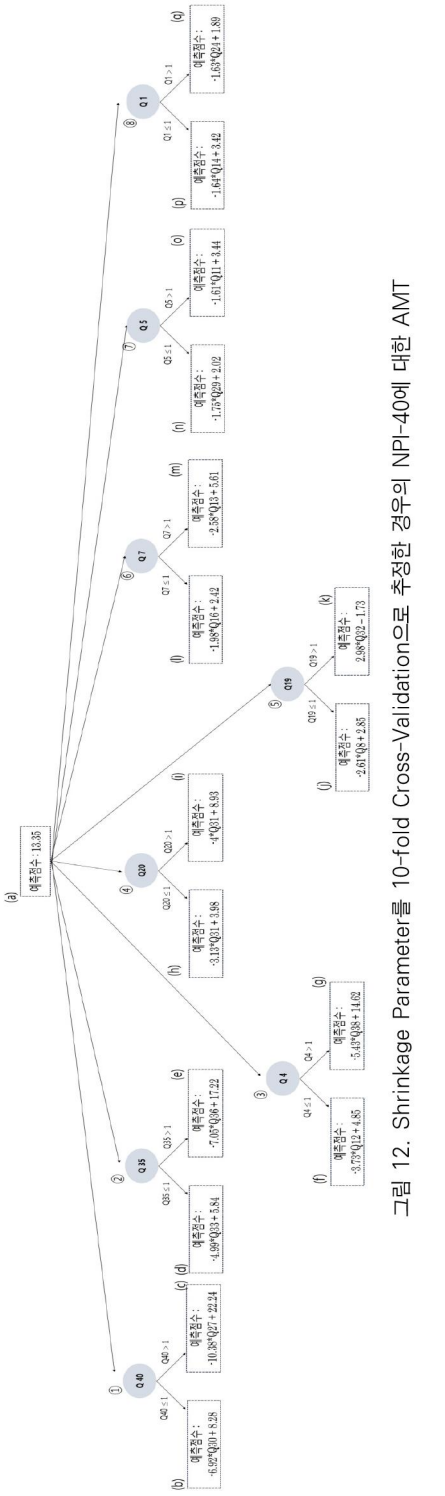

그림 12. Shrinkage Parameter를 10-fold Cross-Validation으로 추정한 경우의 NPI-40에 대한 AMT

표 4. 그림 11에서 가상 피검사자 A, B의 검사과정

	검사과정					
	Decision Stump : 문항	반응		예측노드 : 문항	반응	
가상 피검사자 A	① : Q40	"1"	→	(b) : Q30	"2"	
	② : Q35	"2"	→	(e) : Q36	"2"	
	③ : Q40	"1"	→	(f) : Q27	"1"	
	④ : Q19	"2"	→	(i) : Q27	"1"	
	⑤ : Q8	"1"	→	(j) : Q24	"2"	
	⑥ : Q32	"2"	→	(m) : Q38	"2"	검사종료
가상 피검사자 B	① : Q40	"1"	→	(b) : Q30	"1"	
	② : Q35	"1"	→	(d) : Q33	"2"	→
	⑧ : Q5	"2"	→	(q) : Q13	"2"	
	③ : Q40	"1"	→	(f) : Q27	"2"	
	④ : Q19	"1"	→	(h) : Q27	"2"	
	⑤ : Q8	"2"	→	(k) : Q39	"1"	
	⑥ : Q32	"1"	→	(l) : Q30	"1"	검사종료

표 5. 그림 12에서 가상 피검사자 C, D의 검사과정

	검사과정					
	Decision Stump : 문항	반응		예측노드 : 문항	반응	
가상 피검사자 C	① : Q40	"2"	→	(c) : Q27	"1"	
	② : Q35	"2"	→	(e) : Q36	"1"	
	③ : Q4	"1"	→	(f) : Q12	"2"	
	④ : Q20	"2"	→	(i) : Q31	"1"	
	⑤ : Q19	"1"	→	(j) : Q8	"2"	
	⑥ : Q7	"2"	→	(m) : Q13	"2"	
	⑦ : Q5	"1"	→	(n) : Q29	"2"	
	⑧ : Q1	"2"	→	(q) : Q24	"1"	검사종료
가상 피검사자 D	① : Q40	"1"	→	(b) : Q30	"1"	
	② : Q35	"1"	→	(d) : Q33	"1"	
	③ : Q4	"2"	→	(g) : Q38	"2"	
	④ : Q20	"1"	→	(h) : Q31	"1"	
	⑤ : Q19	"2"	→	(k) : Q32	"2"	
	⑥ : Q7	"1"	→	(l) : Q16	"2"	
	⑦ : Q5	"2"	→	(o) : Q11	"2"	
	⑧ : Q1	"1"	→	(p) : Q14	"1"	검사종료

러므로 이러한 Decision Stump의 분할노드에 존재하는 문항은 피검사자 모두에게 출제되는 공통문항이라고 할 수 있다. 하지만 이에 대한 반응에 따라 각 피검사자는 서로 다른 예측노드에 속할 수 있으며, 이에 따라 출제되는 다음 문항이 서로 다를 수 있었다. 예를 들어, 그림 12에서 피검사자 C와 피검사자 D는 동일한 8개의 Decision Stump를 거쳐 공통문항으로 Q40, Q35, Q4, Q20, Q19, Q7, Q5, Q1를 출제 받았지만, 피검사자 C는 이에 대한 반응으로 Q27, Q36, Q12, Q31, Q8, Q13, Q29, Q24를 다음 문항으로 출제받았다. 반면에 피검사자 D는 공통문항에 대한 반응으로 Q30, Q33, Q38, Q31, Q32, Q16, Q11, Q14를 다음 문항으로 출제받았다. 이와 같은 특징은 그림 11과 그림 12의 AMT 모두에서 관찰되었다. 뿐만 아니라, 그림 11과 같은 형태의 AMT의 경우 앞 문항에 대한 응답에 따라 거치게 되는 Decision Stump가 서로 다를 수 있고, 이에 따라 제시될 수 있는 문항들도 존재했다. 예를 들어, 그림 11에서 예측노드(c)에 속하지 않는 피검사자는 Decision Stump⑦을 거칠 수 없으며, 예측노드(d)에 속하지 않는 피검사자는 Decision Stump⑧을 거칠 수 없다. 결과적으로 검사모형으로서의 AMT로부터 각 피검사자는 서로 다른 검사를 제시 받는다. 그리고 이러한 모든 점은 AMT 구성 과정을 통해 학습된, 검사총점 관측값과 AMT의 예측값 사이의 차이를 최소화하기 위해 선택된 검사과정이다. 그러므로 AMT는 피검사자에게 맞춤형 검사를 제공한다.

두 번째로, 검사모형으로서의 AMT는 높은 검사 시행 효율성을 제공할 수 있다. AMT를 구성할 때 생성할 Decision Stump의 수는 검사 제작자가 직접 입력할 수 있는 파라미터이며, 이는 검사모형으로서의 AMT가 피검사자에게 출제할 수 있는 최대 문항 수를 나타낸다. 예를 들어, 앞서 언급한 것처럼 그림 11과 그림 12의 AMT를 구성할 때 Decision Stump의 수를 8개로 지정했는데 이는 AMT가 최대 16개의 문항을 특정 피검사자에게 출제할 수 있음을 의미한다. 즉, 출제 문항 수를 검사 제작자가 조절할 수 있으며, 이와 동시에 AMT는 항상 문제 은행의 모든 문항에 반응했을 때 얻을 수 있는 검사총점을 예측할 수 있다.

앞 문항에 대한 피검사자의 응답에 따라 가장 큰 정보를 줄 수 있는 다음 문항을 문제은행으로부터 선택해 제시함으로써 각 피검사자별로 맞춤형 검사를 제공하는 것은 컴퓨터 기반 적응적 검사를 정의하는 특징에 해당한다. 또한 높은 검사 시행 효율성과 높은 검사정확도는 컴퓨터 기반 적응적 검사를 적용하는 목적에 해당하는 특징들이다.

정리하자면, 검사모형으로서의 AMT는 컴퓨터 기반 적응적 검사를 정의하는 특징을 가지며, 이를 적용하는 목적인 높은 검사 시행 효율성을 달성할 수 있다. 그러므로 본 논문에서는 AMT에 기반을 둔 AMT-based CAT을 제작 및 사용할 수 있다고 결론 내렸다.

두 번째 연구

두 가지 심리검사에 대한 동일한 학습데이터로 DT-based CAT과 AMT-based CAT을 구성하고 검사정확도를 출제 문항 수를 기준으로 비교한 결과가 표 6과 표 7에 각각 나와있다.

먼저 표 6은 NPI-40에 대한 두 적응적 검사모형의 검사정확도 결과다. 이를 보면, 출제문항 수가 2개인 경우부터 8개인 경우까지는 DT-based CAT이 AMT-based CAT보다 더 높은

표 6. NPI-40에 대한 DT-based CAT과 AMT-based CAT의 검사정확도 결과

출제 문항 수	DT-based CAT			AMT-based CAT		
	상관계수	MAE	RMSE	상관계수	MAE	RMSE
2개	.75	4.50	5.65	.74	4.56	5.73
4개	.85	3.56	4.49	.84	3.62	4.59
6개	.89	3.09	3.90	.88	3.18	3.99
8개	.91	2.80	3.56	.90	2.88	3.66
10개	.92	2.64	3.38	.92	2.70	3.35
12개	.92	2.59	3.34	.93	2.50	3.11
14개	.92	2.56	3.34	.94	2.33	2.92
16개	.92	2.52	3.30	.94	2.20	2.78
18개	.92	2.52	3.32	.95	2.13	2.68
20개	.93	2.50	3.28	.95	2.03	2.56

상관계수, 더 낮은 MAE와 RMSE를 보이는 것으로 나타났다. 하지만 출제 문항 수가 12개인 경우부터는 AMT-based CAT이 일관적으로 더 높은 검사정확도를 보이기 시작했다. 그리고 이 차이는 출제 문항 수가 점점 늘어남으로써 커지는 경향을 보이고 있다. 예를 들어, 출제 문항 수가 12개인 경우에는 상관계수에서 0.01, MAE에서 0.09, RMSE에서 0.23의 차이를 보인 반면에 출제 문항 수가 20개인 경우에는 상관계수에서 0.02, MAE에서 0.47, RMSE에서 0.72의 차이를 보이고 있다.

다음으로 표 7은 EQSQ의 두 하위척도 Empathy Quotient와 Systemizing Quotient에 대한 두 적응적 검사모형의 검사정확도 결과를 보여주고 있다.

첫 번째 하위척도인 Empathy Quotient에 대한 결과를 보면, 출제 문항 수가 2개인 경우부터 6개인 경우까지는 오히려 DT-based CAT이 더 높은 상관계수, 더 낮은 MAE, RMSE를 보였다. 하지만 출제문항 수가 8개인 경우부터는 AMT-based CAT이 더 높은 검사정확도를 보이기 시작했다. 그리고 이러한 차이는 출제 문항 수가 늘어남에 따라 커지고 있다. 또한 DT-based CAT의 경우 14개를 출제한 이후부터는 문항출제를 멈춘 반면 AMT-based CAT은 20개까지 문항출제가 가능한 것으로 나타났다.

계속해서 두 번째 하위척도인 Systemizing Quotient에 대한 결과를 보면 모든 출제 문항 수 상황에서 AMT-based CAT이 DT-based CAT보다 더 높은 상관계수, 더 낮은 MAE, RMSE를 보였다. 그리고 여기서도 출제 문항 수가 늘어남에 따라 검사정확도 차이가 커지는 경향성을 발견할 수 있었다.

정리하자면, 동일한 학습데이터가 주어진 상황에서 출제 문항 수를 기준으로 두 적응적 검사모형의 검사정확도를 비교했을 때 AMT-based CAT이 항상 DT-based CAT보다 뛰어난 검사정확도를 보이지는 못했다. 즉, 출제 문항 수가 적을 때 더 좋은 정확도를 보였지

표 7. EQSQ에 대한 DT-based CAT과 AMT-based CAT의 검사정확도 결과

Empathy Quotient						
출제 문항 수	DT-based CAT			AMT-based CAT		
	상관계수	MAE	RMSE	상관계수	MAE	RMSE
2개	.47	4.13	5.19	.41	4.27	5.36
4개	.58	3.79	4.78	.56	3.88	4.88
6개	.63	3.61	4.57	.63	3.65	4.58
8개	.66	3.54	4.47	.68	3.43	4.33
10개	.67	3.49	4.41	.72	3.25	4.08
12개	.67	3.50	4.45	.74	3.15	4.00
14개	.68	3.46	4.37	.76	3.05	3.84
16개	-	-	-	.78	2.93	3.70
18개	-	-	-	.81	2.73	3.43
20개	-	-	-	.83	2.63	3.30
Systemizing Quotient						
출제 문항 수	DT-based CAT			AMT-based CAT		
	상관계수	MAE	RMSE	상관계수	MAE	RMSE
2개	.74	7.73	9.65	.77	7.46	9.30
4개	.83	6.30	7.95	.84	6.19	7.80
6개	.87	5.64	7.13	.88	5.38	6.71
8개	.89	5.24	6.63	.90	4.93	6.20
10개	.89	5.06	6.50	.92	4.60	5.72
12개	.90	5.00	6.43	.92	4.35	5.49
14개	.90	5.09	6.53	.94	3.98	4.98
16개	.90	5.00	6.44	.94	3.78	4.73
18개	.90	4.99	6.43	.95	3.61	4.55
20개	.90	4.97	6.40	.95	3.46	4.37

만 그 차이가 미미하거나 DT-based CAT이 더 뛰어난 검사정확도를 보이는 경우가 관찰되었다. 하지만 출제 문항 수가 일정 수를 넘어서면 AMT-based CAT이 더 높은 검사정확도를 일관적으로 보이는 경향성을 발견할 수 있었다. 그리고 이러한 검사정확도 차이는 출제 문항 수가 늘어남에 따라 지속적으로 커지는 것으로 나타났다.

논 의

요약

본 논문의 목적은 AMT를 소개하고 AMT-based CAT의 가능성을 탐색하는 데 있다. 이를 위해 먼저 검사 모형으로서의 AMT 작동 방식을 예시를 통해 소개하였고, 다음으로 두 가지 심리 척도 데이터를 대상으로 AMT-based CAT을 구현하고 그 성능을 DT-based CAT과 비교하였다. 연구의 결과는 다음과 같이 요약할 수 있다.

첫째, AMT는 적응적 검사의 특징을 가지는 것으로 확인되었다. 즉, 검사 총점을 추정하기 위해 문항이 순차적으로 제시되며, 이전 문항에 대한 반응에 따라 다음 문항이 선정되는 특징을 가지고 있어 피검사자 별로 맞춤형 검사를 실시할 수 있다. 이는 AMT-based CAT이 가능함을 의미한다.

둘째, AMT-based CAT의 성능은 DT-based CAT의 성능과 유사하거나 좀 더 나은 결과를 보였다. 출제 문항 수가 적은 경우 검사정확도 차이는 미미하거나 DT-based CAT이 약간 더 우수하였다. 그러나, 소수의 문항만을 출제하도록 하는 외부적 제약 상황이 존재하지 않는 한 AMT-based CAT의 성능이 보다 더 나은 것으로 드러났다. 즉, 출제 문항 수가 전체 문항 수의 20%~25% 이상인 경우 AMT-based CAT의 정확도가 일관적으로 더 높았다.

마지막으로, AMT-based CAT은 DT-based CAT 혹은 IRT-based CAT과는 달리 검사 초반 모든 피검사자에게 공통적으로 출제되는 공통 문항을 가지는 특징을 보이는 것으로 나타났다. 이러한 특징은 상황에 따라 AMT-based CAT의 단점이 될 수도 있고 장점이 될 수 있을 것이다. 우선 특정 문항이 지나치게 자주 선정되는 문항 노출 문제에 대한 통제(exposure control)가 어려울 수 있다는 것은 단점이 될 수 있다. 또한 공통 문항은 모든 피검사자에게 제시되기 때문에 검사 보안(test security) 유지가 취약해 질 수도 있다. 그러나 이러한 문항 노출과 검사 보안 관련 이슈들은 AMT-based CAT만의 문제는 아니며, 고위험 검사(high-stakes tests)가 아닌 대부분의 심리 검사 장면에서는 큰 문제가 되지 않을 수 있다. 공통 문항의 존재는 CAT 운영 방식에 따라 장점으로 작용할 수도 있다. 즉, 검사 초기에 공통문항을 묶어 하나의 모듈로 제시하면, 피검사자는 제시된 문항들을 먼저 검토한 후 각자 원하는 순서로 응답을 할 수 있고, 공통 문항 전체에 대한 응답을 최종 제출하기 전 특정 문항에 대한 응답을 수정할 기회도 가질 수 있다. 이후에는 일반적인 CAT과 동일한 절차를 따르게 된다. 이러한 방식은 컴퓨터 기반 다단계 검사(Computerized Multistge Tests; MST; Yan et al., 2016) 방식과 일반적인 CAT이 혼합된 형태로 이해할 수 있다.

한계 및 제안점

본 연구는 AMT-based CAT의 가능성을 확인하였다는 데 가장 큰 의의가 있다. 다만, 선행 연구를 찾기 어려운 연구의 초기 단계이므로 AMT-based CAT의 정확한 평가에 있어서 본 연구에서 사용한 가상 자료와 검사 자료 분석 결과만으로는 한계가 있을 수밖에 없다. 따라서, 추후 보다 다양한 조건(예: 결측치 존재)에서 IRT-based 그리고 DT-based CAT과 그 성능(예: 최적 문항의 수)을 비교하는 체계적인 연구가 필요할 것이다. 특히, 결정-트리는 학

습데이터의 크기가 작을수록 정확도가 낮아진다(Maimon & Rokach, 2014)는 것이 잘 알려져 있으므로, 학습데이터의 크기가 크지 않은 상황에서 AMT-based CAT이 더 높은 정확도를 보일 것으로 예측할 수 있으며, 추후 연구를 통해 이를 확인할 필요가 있을 것이다. 데이터 분포의 특징 및 학습 데이터의 크기가 IRT-based CAT, DT-based CAT 그리고 AMT-based CAT의 검사 정확도에 미치는 영향을 체계적으로 비교 연구하여 예측 성능을 보다 객관적으로 평가할 필요가 있을 것이다.

참고문헌

Antal., M. (2013). On the use of elo rating for adaptive assessment. *Studia Universitatis Babes-Bolyai, Informatica, 58*(1), 29-41.

Babcock, B., & Weiss, D. J. (2009). Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In D. J. Weiss (Eds.), *Proceedings of the 2009 GMAC conference on computerized adaptive testing* (Vol. 14). http://www.psych.umn.edu/psylabs/CATCentral/

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., & Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 358*(1430), 361-374. https://doi.org/10.1098/rstb.2002.1206

Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders, 34*(2), 163-175. https://doi.org/10.1023/b:jadd.0000022607.19833.00

Bjorner, J. B., Chang, C. H., Thissen, D., & Reeve, B. B. (2007). Developing tailored instruments: item banking and computerized adaptive assessment. *Quality of Life Research*, *16*(1), 95-108. https://doi.org/10.1007/s11136-007-9168-6

Breiman, L., J. H. Friedman, R. A. Olshen, & Stone, C. J. (1984). *Classification and Regression Trees.* CRC Press.

Brownlee, J. (2016). *Master Machine Learning Algorithms: discover how they work and implement them from scratch.* Machine Learning Mastery.

Carlo, A. D., Barnett, B. S., & Cella, D. (2021). Computerized Adaptive Testing (CAT) and the Future of Measurement-Based Mental Health Care. *Administration and Policy in Mental Health and Mental Health Services Research*, *48*, 729-731. https://doi.org/10.1007/s10488-021-01123-9

Delgado-Gomez, D., Baca-Garcia, E., Aguado, D., Courtet, P., & Lopez-Castroman, J. (2016). Computerized adaptive test vs. decision trees: development of a support decision system to identify suicidal behavior. *Journal of affective disorders*, *206,* 204-209. https://doi.org/10.1016/j.jad.2016.07.032

Delgado-Gómez, D., Laria, J. C., & Ruiz-Hernández, D. (2019). Computerized adaptive test and decision trees: A unifying approach. *Expert Systems with Applications*, *117,* 358-366.

https://doi.org/10.1016/j.eswa.2018.09.052

Dietterich, T. G., & Kong, E. B. (1995). *Machine learning bias, statistical bias, and statistical variance of decision tree algorithms*. Department of Computer Science, Oregon State University.

Domingos, P. (2000). A unified bias-variance decomposition. In P. Langley (Eds.), *Proceedings of 17th International Conference on Machine Learning* (pp. 231-238). Morgan Kaufmann.

Frank, E., Mayo, M., & Kramer, S. (2015). Alternating model trees. In R. L. WainWright (Eds.), *Proceedings of the 30th annual ACM symposium on applied computing* (pp. 871-878). ACM.
https://doi.org/10.1145/2695664.2695848

Gibbons, R. D., Hooker, G., Finkelman, M. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T. & Kupfer, D. J. (2013). The computerized adaptive diagnostic test for major depressive disorder (CAD-MDD): a screening tool for depression. *The Journal of clinical psychiatry, 74*(7), 669-674.
https://doi.org/10.4088/jcp.12m08338

Gibbons, R. D., Weiss, D. J., Frank, E., & Kupfer, D. (2016). Computerized adaptive diagnosis and testing of mental health disorders. *Annual review of clinical psychology*, *12,* 83-104.
https://doi.org/10.1146/annurev-clinpsy-021815-093634

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning.* Springer. https://doi.org/10.1007/b94608

Hsueh, I. P., Chen, J. H., Wang, C. H., Chen, C. T., Sheu, C. F., Wang, W. C., Hou, W. H., & Hsieh, C. L. (2010). Development of a computerized adaptive test for assessing balance function in patients with stroke. *Physical therapy, 90*(9), 1336-1344.
https://doi.org/10.2522/ptj.20090395

Hu, Y., Cai, Y., Tu, D., Guo, Y., & Liu, S. (2020). Development of a Computerized Adaptive Test for Separation Anxiety Disorder Among Adolescents. *Frontiers in Psychology*, *11,* 1077.
https://doi.org/10.3389/fpsyg.2020.01077

Istiyono, E., Dwandaru, W. S. B., Setiawan, R., & Megawati, I. (2020). Developing of Computerized Adaptive Testing to Measure Physics Higher Order Thinking Skills of Senior High School Students and Its Feasibility of Use. *European Journal of Educational Research, 9*(1), 91-101.
https://doi.org/10.12973/eu-jer.9.1.91

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning.* Springer.
https://doi.org/10.1007/978-1-4614-7138-7

Kimura, T. (2017). The impacts of computer adaptive testing from a variety of perspectives. *Journal of educational evaluation for health professions*, *14*(12), 1-5.
https://doi.org/10.3352/jeehp.2017.14.12

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.
https://doi.org/10.1007/978-1-4614-6849-3

Lai, M. C., Lombardo, M. V., Chakrabarti, B., Ecker, C., Sadek, S. A., Wheelwright, S. J., Murphy, D. G. M., Suckling, J., Bullmore, E. T., MRC AIMS Consortium & Baron-Cohen,

S. (2012). Individual differences in brain structure underpin empathizing-systemizing cognitive styles in male adults. *Neuroimage*, *61*(4), 1347-1354. https://doi.org/10.1016/j.neuroimage.2012.03.018

Linden, W. J., van der Linden, W. J., & Glas, C. A. (Eds.). (2000). *Computerized adaptive testing: Theory and practice*. Springer. https://doi.org/10.1007/0-306-47531-6

Ma, X. (2018). *Using classification and regression trees: A practical primer*. IAP.

Magis, D., Yan, D., & Von Davier, A. A.(2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer. https://doi.org/10.1007/978-3-319-69218-0

Maimon, O. Z., & Rokach, L. (2014). *Data mining with decision trees: theory and applications* (Vol.81). World scientific.

Michel, P., Baumstarck, K., Loundou, A., Ghattas, B., Auquier, P., & Boyer, L. (2018). Computerized adaptive testing with decision regression trees: an alternative to item response theory for quality of life measurement in multiple sclerosis. *Patient preference and adherence, 12,* 1043. https://doi.org/10.2147/ppa.s162206

Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*(1), 59-71. https://doi.org/10.1177/014662169001400106

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics, 18*(6), 275-285. https://doi.org/10.1002/cem.873

Peute, L., Scheeve, T., & Jaspers, M. (2020). Classification and Regression Tree and Computer Adaptive Testing in Cardiac Rehabilitation: Instrument Validation Study. *Journal of medical Internet research, 22*(1), e12509. https://doi.org/10.2196/preprints.12509

Raskin, R. N., & Hall, C. S. (1979). A narcissistic personality inventory. *Psychological Reports*, *45,* 590.

Raskin, R., & Terry, H. (1988). A principal-components analysis of the Narcissistic Personality Inventory and further evidence of its construct validity. *Journal of personality and social psychology*, *54*(5), 890. https://doi.org/10.1037/0022-3514.54.5.890

Rezaie, M., & Golshan, M. (2015). Computer adaptive test (CAT): Advantages and limitations. *International Journal of Educational Investigations, 2*(5), 128-137.

Seo, D. G., Lee, S. M., Kim, J. N., Choi, S. W., Chae, J. M., Jung, S. H., Cho, S. K., Kim, M. K. & Ebesutani, C. (2019). Psychometric Methods and Validation of Short Form for the Psychological Scale: Based on the Korean dysfunctional depression scale. *Korean Journal of Psychology: General 38*(1), 75-102. http://dx.doi.org/10.22257/kjp.2019.3.38.1.75

Šerbec, I. N., Žerovnik, A., & Rugelj, J.(2011). Adaptive assessment based on decision trees and decision rules. In A. Verbraeck, M. Helfert, J. Cordeiro, & B. Shishkov (Eds.), *CSEDU 2011-Proceedings of the 3rd International Conference on Computer Supported Education* (pp. 473-479). SciTePress. https://doi.org/10.5220/0003521104730479

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130. https://doi.org/10.11919/j.issn.1002-0829.215044

Sug, H. (2009). An effective sampling method for decision trees considering comprehensibility and accuracy. *WSEAS Transactions on Computers*, *8*(4), 631-640.

Thissen, D. E., & Wainer, H. E. (2001). *Test scoring*. Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410604729

Twenge, J. M., Konrath, S., Foster, J. D., Keith Campbell, W., & Bushman, B. J. (2008). Egos inflating over time: A cross temporal meta analysis of the Narcissistic Personality Inventory. *Journal of personality*, *76*(4), 875-902. https://doi.org/10.1111/j.1467-6494.2008.00507.x

Ueno, M., & Songmuang, P. (2010). Computerized adaptive testing based on decision tree. In B. Werner (Eds.), *2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 191-193). IEEE Computer Society Press. https://doi.org/10.1109/icalt.2010.58

Van der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Elements of adaptive testing* (pp. 3-30). Springer. https://doi.org/10.1007/978-0-387-85461-8_1

van der Oest, M. J., Porsius, J. T., MacDermid, J. C., Slijper, H. P., & Selles, R. W. (2020). Item reduction of the patient-rated wrist evaluation using decision tree modelling. *Disability and rehabilitation, 42*(19), 2758-2765. https://doi.org/10.1080/09638288.2019.1566407

Walter, O. B., Becker, J., Bjorner, J. B., Fliege, H., Klapp, B. F., & Rose, M. (2007). Development and evaluation of a computer adaptive test for 'Anxiety'(Anxiety-CAT). *Quality of Life Research*, *16*(1), 143-155. https://doi.org/10.1007/s11136-007-9191-7

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge. https://doi.org/10.4324/9781410605931

Winarno, D., & Si, S. (2018). Computerized Adaptive Testing (CAT) Using Triangle Decision Tree Method. *International Journal of Science and Research*, 7(5), 552-560. https://doi.org/10.21275/ART20182213

Wu, C. C., Chen, Y. L., Liu, Y. H., & Yang, X. Y. (2016). Decision tree induction with a constrained number of leaf nodes. *Applied Intelligence*, *45*(3), 673-685. https://doi.org/10.1007/s10489-016-0785-z

Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics*, *29*(3), 293-316. https://doi.org/10.3102/10769986029003293

Yan, D., Von Davier, A. A., & Lewis, C. (Eds.). (2016). *Computerized multistage testing: Theory and applications*. CRC Press.

1차원고접수 : 2021. 08. 24.
2차원고접수 : 2021. 10. 10.
최종게재결정 : 2021. 10 .25.

한국심리학회지: 일반
Korean Journal of Psychology: General
2021, Vol. 40, No. 4, 539-566
http://dx.doi.org/10.22257/kjp.2021.12.40.4.539

Investigating the Viability of Alternating Model Tree As An Item Selection Algorithm for Constructing Computerized Adaptive Psychological Testing

Jeong-Han Youn **Taehun Lee**

Department of Psychology, Chung-Ang University

Computerized adaptive testing (CAT) is a computer-administered test where the next question for estimating the examinee's trait level is selected depending on his or her reponses to the previous items, resulting in tailored testing for each individual examinee. A defining feature of CAT stems from its item selection algorithms, among which both research interest and practical applications of decision-tree based CAT (DT-based CAT) have been rising recently. In the field of machine learning, however, it is well known that decision-trees, as a form of predictive models with simple and interpretable tree structures, can be vulnerable to the problem of overfitting or the problem of creating overly complex trees that do not generalize to newly observed data. Among various ensemble techniques developed to adequately address this problem, we the authors paid attention to the Alternating Model Tree (AMT) due to its interpretable tree-like structure. The purpose of this article is to investigate the viability of the Alternating Model Tree (AMT) as an item selection algorithm for constructing CAT. To this end, we first presented a detailed exposition of how AMT-based CAT can be constructed and then compared its performance with DT-based CAT using two sets of publicly available psychological test scores. The results provided supportive evidence that AMT-based CAT is viable, and that AMT-based CAT can predict test scores at least as accurate as DT-based CAT does. Based on our findings, we discuss implications, limitations, and directions of future studies.

Key words : Computerized Adaptive Test, Decision Tree, Alternating Model Tree, Item Selection Algorithm