SPRT와 EXSPRT-R 검증법의 언어능력 시험적용에 대한 연구

김명관*, 김지한*
*서울보건대학 컴퓨터정보과
e-mail:binsum@sh.ac.kr

A Study of SPRT and EXSPRT-R Appling Foreign Language Test

Myung-Gwan Kim*, Ji-Han Kim*
*Dept of Computer Information, Seoul Health College

요 약

CAT(Computer Adaptive Testing : 컴퓨터 기반 적응적 검사)는 기존의 종이 시험지에서 이루어지던 시험과 달리 수험자에게 적절한 맞춤식 출제로 보다 정확한 수험자의 능력 판단 및 빠른 수험진행을 가능케 하였다. 기존의 CAT는 많은 인원과 문제가 있어야만 그 결과에 신뢰성이 있다고 알려져 있다. CAT의 대표적인 알고리즘인 SPRT와 EXSPRT-R을 이용하여 10명의 적은 인원으로 JLPT 4급 기출문제를 적용한 실험을 하였다. SPRT 에서는 인원수와, 문제 난이도를 무시한 결과로 인하여 만족 할만한 결과를 얻지 못하였으나, EXSPRT-R의 경우에는 적은 인원에서도 충분히 CAT를 이용할 수 있음을 발견할 수 있었다.

1. 서론

현대인에게 있어서 개인의 능력을 평가하는 각종 시험들은 그들을 심사할 수많은 평가자를 대신하여 검증해 주는 객관적 척도의 대리 평가의 역할을 하고 있다. 그러나 기존의 필기로 진행되는 지면을 이용한 시험을 통해서는 공신력이 있는 기관에서 주관하는 시험이라 할지라도 수험자의 기량에 대하여 자격증에 대한 명확한 신뢰를 둘 수 없는 것이 현실이며, 이에 대한 해결책이 오랫동안 요구되어 왔다. 그로 인해 등장한 컴퓨터 기반 적응적 검사(CAT : Computerized Adaptive Testing)는 수험시간의 낭비와, 문제의 노출 위험이 높은 기존의 시험과 달리 각종 시험 시 수험자의 성적 수준을 결정하는데 필요한 문항을 최소화 하고 수험자 개인의 수준에 맞추어 각각 다른 상황에 맞추어 문제를 제시함으로써 보다 빠르고, 효율적인 수험자의 기량 파악이 가능해질 수 있다[1,2]. 일반적, 제한적인 기존의 평가 방식 내에서는 수험자의 취약점에 대해서 관련된 반복적 질문 시 심도 있는 추적에 한계가 있으므로, 이를 극복하여 좀 더 구체적이고 확실한 수험자에 대한 평가를 거두는 것이 CAT의 목적이다[3].

1990년대 중반 이전까지만 하더라도 CAT형식의 시험은 플로피 디스크를 통하여 실시되는 것이 대부분이었지만, 90년대부터는 고용량의 CD-ROM을 이용하기 시작하였고, 최근에는 인터넷등의 네트워크를 통한 수검이 보편화되기 시작하였다[4]. 대표적인 CAT 알고리즘의 적용 예는 TOFEL(Testing of English as a Foreign Language)의 CBT(Computer Based TOFEL)이다. 기존의 종이 시험지 상에서 이루어지던 PBT(Paper Based TOFEL)를 대신하여 국내에서는 2000년 9월부터 시행되기 시작한 새로운 TOFEL 시험의 방식으로서, 컴퓨터를 통하여 수험자가 시험을 보게 되는 CAT알고리즘을 적용한 방식으로 시험의 응시 시기와 응시 장소의 선택이 용이할 뿐만 아니라 초반 문제를 이용한 수검자의 능력 평가를 이용하여 이후 제시되는 문제들의 난이도를 조정, 시험자의 수준에 맞게 적절히 문제를 제시하며, 수험자에 대한 평가를 시험이 끝나는 동시에

확인할 수 있는 장점이 있다[5]. 수험자의 정확한 능력에 대한 판단이 중요한 어학시험에서 PBT와 같은 기존의 지면시험은 평균적 능력을 갖고 있는 학생을 파악하기에는 용이할지라도 시험에 지원한 학생들의 광범위한 능력분포까지는 파악하기 힘들었다. 그러나 CBT와 같은 CAT알고리즘을 적용할 시에는 보다 객관적이고 구체적인 수험자의 파악을 가능하게 할 수 있다[6].

본 실험에서 기존의 지면상의 필기시험과 CAT의 대표적인 기법인 SPRT와 EXSPRT-R과의 비교를 통하여 지면을 이용한 필기시험에 대한 CAT가 갖고 있는 차이점과 장, 단점을 분석하였다. 또한 많은 인원의 수험자를 요구하는 CAT의 기존의 개념을 상대적으로 적은 수의 수험자에 적용 하여 그 신빙성에 대하여 조사한 결과 적은 인원과 문항이 주어진 환경에서도 충분히 CAT를 적용할 수 있음을 확인하였다.

2. 실험의 준비 및 개념

CAT알고리즘을 이용한 대표적인 검증법으로는 SPRT(Sequential Probability Ratio Testing)와 S-PRT의 단점을 보완한 EXSPRT-R(Expert Sequential Probability Ratio Testing-Random)이 있다.

실험에 앞서 이 두 알고리즘에 대한 기초적인 수행과정에 대해 알아보고자 한다.

2.1. SPRT(순차적 확률비 검증법)

구조가 단순하면서 결과 오차가 정확하여 쉽게 사용할 수 있는 검증법이다.

예를 들어 10문제의 시험을 시행 시 합격자와 불합격자가 발생할 경우, 한 문제를 풀 때마다 정해진 SPRT 규칙을 이용하여 어떠한 문제를 맞게 답할 경우 합격자일 경우와 틀렸을 경우 불합격자일 경우를 예상한 '사후 확률비(Posterior probability ratio)'를 이용하여 오류비(α, β)에 비교하여 합격 및 불합격 여부를 결정 후 시험을 계속 진행할지 혹은 종료할지의 여부를 결정한다.

사후 확률비는 (수식 1)과 같다.

$$PR = \frac{(PomPm)^r(1-Pm)^w}{(PonPn)^r(1-Pn)^w}$$

(수식 1) SPRT의 사후확률비 계산

PR - 사후확률비
Pom - 합격의 사전 확률치를 나타내는 초기값
Pon - 불합격의 사전 확률치를 나타내는 초기값
Pm - 합격이라는 가설 하에서 정답 검사문항을 선택할 확률
Pn - 불합격이라는 가설 하에서 정답 검사문항을 선택할 확률
r - 관찰된 정답의 수
w - 관찰된 오답의 수

α는 주어진 첫 번째 판단이 진실로 참인데도 거짓이라고 잘못 판단할 확률이며, β는 이와 반대의 경우이다.

오류의 허용범위를 나타내는 특정의 신뢰수준 하에서 두가지 판단 중의 하나로 결정될 때 수험의 합격 혹은 불합격 여부나 종료 여부의 판단의 기준은 $PR \geq (1-\beta)/\alpha$ 시 더 이상의 관찰을 중지하고 '합격'이라는 가설을 채택하고, $PR \leq \beta(1-\alpha)$이면, 더 이상의 관찰을 중지하고 '불합격'이라는 가설을 채택한다. 만약 $\beta/(1-\alpha) < PR < (1-\beta)/\alpha$ 이면 다른 한문항을 무작위 선택하고 학생들에게 문항 풀이를 계속 실시하여 합격이나, 불합격의 가설에 부합할 때까지 이 과정을 계속한다[7].

2.2. EXSPRT-R

SPRT는 시험의 각 문항의 난이도가 다양할 시 수험자의 성취수준을 평가하기에 부적절한 단점이 있다. 즉, 합격일 수험자가 어떤 문항에서 정답을 택할 확률이 타 문항에서 정답을 택할 확률과 다를 수가 있는 경우에서 SPRT의 단점을 발견할 수 있다.

EXSPRT-R은 문항 풀에서 문항을 불러올 시 SPRT와 유사하게 작동하나, 합격자와 불합격자를 판정하는 공식의 적용에 있어서 문항의 난이도를 고려한다는 차이점이 있다. EXSPRT-R의 사후확률계산은 (수식 2), (수식 3)과 같다.

$$LR = \frac{L1 = Pom^* P(C \mid M)}{L2 = Pon^* P(C \mid N)}$$

(수식 2) EXSPRT-R의 정답 시의 사후확률비 계산

$$LR = \frac{L1 = Pom^* P(\sim C \mid M)}{L2 = Pon^* P(\sim C \mid N)}$$

(수식 3) EXSPRT-R의 오답 시의 사후확률비 계산

합격, 불합격 경우의 사전확률치 값은 현재 테스

트의 상황에 맞게 임의의 값으로 지정되고 계산 후의 L1, L2값은 각각 다음 단계의 사전확률이 된다. 이후 LR > (1-α) / β 이면 '합격'이라는 판정을 내리고 관찰을 중지하고, LR < α/(1-β)이면 '불합격'이라는 판정을 내리고 검사를 종료한다. 마지막으로 α/(1-β) ≤ (1-α)/β 이면 시행을 계속하여 합격이나 불합격의 판정에 부합될 때까지 이 과정을 계속 한다[7].

2.3. 실험의 방법

본 실험에서 이용한 테스트는 국제교류기금(國際交流基金)과 일본국제교육지원협회(日本國際教育支援協會)에서 주관하는 JLPT(Japanese Language Proficiency Test, 일본어능력시험)의 4급 기출문제를 이용하였다. JLPT는 일본어를 모국어로 하지 않는 사람들을 대상으로 1984년부터 시행되고 있는 가장 권위 있는 일본어 시험으로서, 1급에서 4급까지 4단계로 나누어져 있으며, 매년 12월에 시행된다.

JLPT는 '문자・어휘(文字・語彙)'와 '독해・문법(讀解・文法)', 그리고 '청해(聽解,듣기문제)'의 세과목으로 이루어져 있다[8].

본 실험에서 사용한 지문들은 JLPT '문자・어휘'와 '독해・문법'과목의 2000년도 4급 기출 문제를 추려낸 것으로서 실제 JLPT 시험에서의 한 과목인 '청해'의 경우에는 자료수집의 어려움으로 제외시켰다. 실험대상이 될 수험자들은 최근 일본어 어학연수 3개월 코스를 마치고 돌아온 대학생 10명을 대상으로 하였다. 실제 합격 기준은 '문자・어휘'가 100점, '청해' 100점, 그리고 '독해・문법'이 200점 만점으로 총 400점 만점의 60%인 240점이지만 실험 대상으로 한 수검자들의 수준과 통계적 확률을 고려하여 문제의 배점을 변형시켜 '문자・어휘'와 '독해・문법' 두과목의 87문제 중 과반수인 44개 이상의 문항을 맞출 시 합격으로 커트라인을 조종하였다.

우선 10명의 실험자에게 위에 언급한 시험문제를 출제하여 기존 지면상의 시험과 동일한 75분의 제한시간과 주변 환경에서 수험을 거친 뒤 EXSPRT-R에 적용을 하여 사용할 수 있도록 온라인 문제의 데이터베이스를 작성 하여 지면시험 시 각 문제의 정답 및 오답자들의 정보를 입력하였다.

이후 지면상의 시험지를 검토하여, SPRT와 EXSPRT-R 두가지의 검증법을 대입하여 기존 지면상의 테스트와의 결과 및 수험 시간등을 각각 비교 조사하였다.

2.4. 지면을 사용한 기존형태에서의 수험 결과

일반적인 지면을 통한 수험자 10명을 정리한 데이터로서 총 문항 87개중 틀린 문항의 수를 (표 1) 정리하였다.

수험자	수험시간	오답 수	정답 수	합격여부
A01	75분	37개	50개	합격
A02	75분	36개	51개	합격
A03	75분	41개	46개	합격
A04	75분	49개	38개	불합격
A05	75분	64개	23개	불합격
A06	75분	41개	46개	합격
A07	75분	49개	38개	불합격
A08	75분	41개	46개	합격
A09	75분	49개	38개	불합격
A10	75분	44개	43개	불합격

(표 1) 지면을 이용한 시험상의 결과

2.5. SPRT 검증법을 통한 지면 시험지와의 비교

SPRT를 적용하여 보다 객관적인 결과를 기대하기에는 시험 문항의 수가 적었고, 시험 문제의 다양한 난이도를 고려하지 않은 채 무작위로 문제가 제시되는 SPRT의 약점 때문에 검증 시 지면시험지와의 비교에서는 결과에 있어서 상당한 차이를 보였다. (표 2)와 같이 합격과 불합격의 경우가 바뀌는 심각한 오류를 나타내기도 하였으며, 모든 문제를 풀었음에도 불구하고 사후 확률비가 합격 혹은 불합격 양쪽의 수치모두에 못 미쳐 합격여부 판단이 불가능한 경우도 4회 발생하였다.

수험자	지면시험시의 합격여부	SPRT를 통한 합격여부 판단(소요 횟수)
A01	합격	불합격(3회)
A02	합격	합격(30회)
A03	합격	합격(5회)
A04	불합격	판단 불가(87회)
A05	불합격	불합격(3회)
A06	합격	합격(16회)
A07	불합격	판단 불가(87회)
A08	합격	판단 불가(87회)
A09	불합격	판단 불가(87회)
A10	불합격	합격(5회)

(표 2) 지면 시험 결과의 SPRT를 통한 분석

2.6. EXSPRT-R 검증법을 통한 비교

난이도에 따라 문제에 차등을 두는 EXSPRT-R

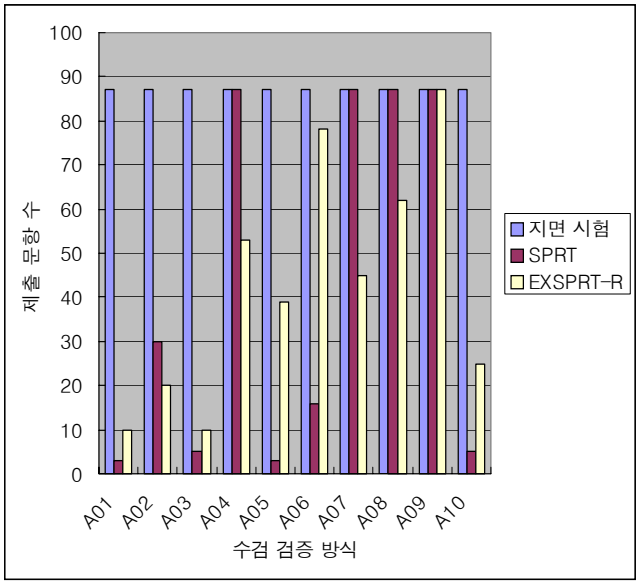
을 이용한 검증 실험에서는 수험자 'A9'의 경우를 제외하고는 (표 3)의 결과와 같이 지면 시험상의 합격과 불합격의 결과에 동일한 결론을 내릴 수 있었다.

수험자	지면시험시의 합격여부	EXSPRT-R을 통한 합격여부 판단(소요 횟수)
A01	합격	합격(10회)
A02	합격	합격(20회)
A03	합격	합격(10회)
A04	불합격	불합격(53회)
A05	불합격	불합격(39회)
A06	합격	합격(78회)
A07	불합격	불합격(45회)
A08	합격	합격(62회)
A09	불합격	판단 불가(87회)
A10	불합격	불합격(25회)

(표 3) 지면 시험 결과의 EXSPRT-R을 통한 분석

2.7. 세가지 검증법의 수검문항 수 비교

CAT 알고리즘을 이용한 검증법의 가장 큰 장점 중 하나로 꼽히는 것이 수검시간의 단축이다. 그러나 지면을 통한 시험 후 SPRT와 EXSPRT-R의 경우에서는 지면시험 결과 자료의 계산을 통하여 분석한 결과물이었던 관계로 수검시간의 잣대가 되어줄 수 있는 각각의 검증법에 대한 출제 된 문항의 수와 그 평균을 각각 (그림 1)과 (표 4)와 같이 그래프와 표로 정리하였다.



(그림 1) 각각의 검증법에 대한 제출 문항 수

수험 검증법	문항수의 평균	합격자 판정의 지면 시험과의 일치 수준
지면 시험	87회	100%
SPRT	41회	40%
EXSPRT-R	42.9회	90%

(표 4) 출제 문항 수의 평균

3. 실험의 결과

SPRT에서는 다양한 문제의 난이도를 무시하고 무작위로 문항을 제시하는 까닭에 다양한 난이도의 문제들이 수록 되었던 본 실험에는 합격자를 불합격자로, 혹은 그 반대의 경우의 결과가 출력되는 심각한 오류를 찾을 수 있었다. 그러나 EXSPRT-R에서는 SPRT보다 훨씬 신뢰할만한 정보를 얻을 수 있었다. 비록, 경우에 따라서는 무작위로 제시되는 문항 출제 순서 때문에 합격 혹은 불합격 여부를 알아내기 위하여 이용한 문항수가 SPRT보다 무척 길었지만, 지면 시험과의 일치 수준이 40%에 머물었던 SPRT의 경우와는 달리 90%의 일치를 보였다.

위 결과는 적은 인원에서도 충분히 EXSPRT-R을 이용한 검증이 가능하다는 것을 말해준다.

참고문헌

[1] 김규봉 외 "SPRT를 이용한 웹 기반 학습자 주도적 학습 시스템" 제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호

[2] T.J.H.M.Eggen "Overexposure and underexposure of items in computerized adaptive testing" http://download.citogroep.nl/pub/pok/reports/Report01-01.pdf

[3] J. Christine Harmes "Item Exposure Control in Computer-Adaptive Testing" http://www.coedu.usf.edu/fjer/2000/FJERV40P2852.pdf

[4] 손미 외 "원격교육의 이론과 실제" 학지사

[5] Bruce Rogers "TOFEL CBT Success 2004" 7th Ed. Petersons Guides

[6] Susan Grist "Computerized Adaptive Tests" ERIC Clearinghouse on Tests Measurement and Evaluation Washington DC. American Institutes for Research Washington DC.

[7] 김영환 외 "컴퓨터 기반 적응적 검사 (CAT)의 이론과 실제" 문음사

[8] "일본어능력시험 공식 홈페이지 JLPT 소개" http://www.jlpt.or.kr/submenu1-5.asp