電気通信大学大学院博士前期課程

平成30年度 情報理工学研究科修士論文

アダプティブテスティングにおける様々な情報量の比較

電気通信大学大学院 情報理工学研究科

情報・ネットワーク工学専攻 情報数理工学プログラム

学籍番号　1631136

増田隆太

主任指導教員　植野 真臣　教授

指導教員　川野 秀一　准教授

2019年1月28日

目 次

第1章　はじめに

1.1　研究の背景

1.1.1　e テスティングの概要

本研究はコンピュータ適応型テスト（Computer-adaptive Testing: CAT）を対象としたものである。CAT とはコンピュータの計算能力を用いて、現在テストを受検している受検者の能力に合わせて、適応的に問題項目（アイテム）を逐次出題することのできるテスティング技術である。コンピュータ上で実施するテストは CBT（Computer-based Testing）と総称され、これらの技術は e テスティングとして近年実用化されている。

e テスティングは能力が自動的かつ高精度に推定できることに加え、その上で異なる項目群に回答した受検者間の能力を同一尺度上で比較できるといった利点を有する。多くの検定試験などでは毎回のテストの難易度および信頼性を標準化しなければならないが、それはテスト構成者の経験や勘のみに頼るもので実現が難しく、その問題を解決するための手法として e テスティングが有効である。受検者にとっても時間と場所を問わず Web 上で受けることができる点から、e テスティングによってテスト構成者と受検者双方の負担が大幅に軽減されるといえる。テストの規模が大きくなればなるほどこれらの利点は重要なものとなり、特にテストの受検者に及ぼす影響の大きいハイステークステストにおいて導入が進んでいる [1]。

e テスティングの代表的な実用化例としては、Test of English as a Foreign Language（TOEFL）が挙げられる。TOEFL は 1998 年にコンピュータベース版が、2005 年にはより幅広いフォーマットの項目を使用できるようインターネットベース版が導入されている [2]。また、国内では情報処理技術者試験の一区分である国家試験の IT パスポート試験 [3] や、医療系大学間共用試験実施評価機構による臨床実習開始前の共用試験 [4] で利用されている [5][6]。

1.1.2 CAT の概要

受検者がテスト中のある項目に対して誤答した場合はさらに難度の低いものを、正答した場合はさらに難度の高いものをリアルタイムで提示することが可能である。CAT は以下のメリットを有する。

- 適応的な出題によって、受検者の測定精度を減少させることなく出題項目数や受検時間を軽減できる
- テスト終了基準を適切に設定することで、全ての受検者を同程度の精度で測定できる [7]

1.1.3 研究の動機

CAT で項目を選択する際には項目の「困難度値」や「識別力値」、また受検者の現在の推定能力値といったパラメータを参照し、ある「情報量」という項目選択基準に基づいて、その値が最も大きくなるような項目を出題する手法を取ることが一般的である。

より受検者に対する能力推定精度が高い情報量を用いることで、提示項目数を少なくすることができる。しかし、能力推定精度が高い情報量は積分を含むなど計算式が複雑であるために計算量が大きく、受検者に対する項目提示が遅れることで負担を強いる可能性がある。計算量の大きさと能力推定精度はトレードオフの関係にある。

Ueno & Songmuang(2010)[8] によって、適応型テストにおいて高速で正確な項目選択アルゴリズムを実現するために、ID3 アルゴリズムによってすべての可能な受験者の応答パターンに対する項目決定木を事前に生成する手法が提案されている。適応テストにおいて高速で正確な項目選択アルゴリズムを実現するために、最近ではこの研究を基にした応用研究が進められている [9][10]。

また Ueno(2013)[11] によって、Lord & Novick(1968)[12] によって提案され一般的に項目選択基準として用いられるフィッシャー情報量（Fisher Information: FI）と比較して、

- 項目選択の偏りが小さい

- テスト開始直後の推定誤差が小さく、それにより推定値の収束が早期かつ高精度である

といった点が長所である情報量 Expected Value of Test Information(EVTI) が提案されている。第 3 章で詳細に説明するが、FI の他にも 3 つの情報量と比較されており、Chang & Ying(1996)[13] の提案するカルバック・ライブラー情報量を含んだ Maximum Global-Information (MGI)、 Veerkamp & Berger(1997)[14] の提案する尤度重み付け関数 Likelihood-Weighted Information (LWI)、Linden(1998)[15] の提案する尤度関数と事前分布を含む Maximum Expected Posterior Weighted-Information (MEPWI)、が対象になっている。それらと比較した結果、EVTI の精度が最も高く、項目選択の偏りが小さく、提示項目数が少ないという結論が出ている。

推定精度の高い情報量は長い計算時間を必要とするが、テストの事前に決定木を生成しておくことによってテスト実施中の項目選択が瞬時に行われるようになる。つまり、候補となる項目が多い場合にも項目選択時間を気にすることなく、決定木生成時の情報量には精度の高いものを用いることができる。推定精度が高くなることによって、CAT の提示項目数が少なくなる。CAT において推定精度の低下なく提示項目数を少なくできることによって、以下のメリットがある。

- テストの回答数が減り時間が短くなるため、テスト受検者の負担が小さくなる
- データベース上の項目の曝露率が小さくなり、同じ項目を長期間使うことができる

決定木を生成する場合にはテストの実施時間には悪影響を及ぼさないとはいえ、EVTI の計算は非常に大きく、高精度なテストを構成しようとするほどに現実的な時間で生成することが難しくなる。そこで、本研究では EVTI を用いた CAT を効率化するアルゴリズムを提案する。

また、以上に述べた情報量の相対的な精度や提示項目数が変化する条件が発見されれば、テストの特徴によって項目選択基準としての情報量のより有効な採用指針を考えることができる。本研究では EVTI とその他情報量のさらなる特性を明らかにすべく、以下の条件でシミュレーション実験を行う。

- 通常の項目群を用いる

- 識別力が低い項目群を用いる
- 被検者の能力真値から外れた困難度をもつ項目群を用いる

1.2 研究の目的

先行研究 [11] において、計算量が大きくなるという欠点があるものの能力値の推定精度が比較的高く、その他の情報量を用いた場合よりも信頼性の高いテスティングを可能にすると考えられている EVTI という情報量（第 4 章で説明する）が提案されている。本研究では、EVTI を用いた場合における CAT アルゴリズムの効率化手法を提案する。また、通常の CAT に条件を付加してシミュレーション実験を行いその情報量の特性を明らかにする。それらによる成果から、現在の e テスティングシステムに対して全体のテスト時間が長くなるなど快適性を損なうようなことなく高精度なテストを提供できるようにすることが目的である。

1.3 本論文の構成

本論文は、本章を含めて 8 章から構成される。第 2 章では、コンピュータ適応型テストにおいて重要な理論である項目反応理論について説明する。第 3 章では、項目反応理論を用いたコンピュータ適応型テストについて説明する。第 4 章では、高精度な予測情報量である Expected Value of Test Information および、提案手法としてそれを用いた適応型テストの効率化アルゴリズムについて説明する。第 5 章では、本実験で行うシミュレーション実験の設定を述べる。第 6 章では、第 5 章で説明した実験の結果を表とグラフで示し、実験結果についての考察を述べる。最後に第 7 章では、本論文で記述する研究内容およびその結果を総括する。

第2章　項目反応理論

2.1　2パラメータロジスティックモデル

項目反応理論（Item Response Theory: IRT）とは、テストに用いられる各項目の困難度と識別力および受検者の能力等を、受検者の回答パターンから評価するためのテスト理論である。本研究では現在広く用いられている、受検者 j が項目 i に正答する確率を式 (2.1) で表す 2 パラメータロジスティックモデル（2PLM）を採用する。

$$p(u_{ij}=1|\theta_j)=\frac{1}{1+\exp(-Da_i(\theta_j-b_i))} \tag{2.1}$$

u_{ij} は式 (4.1) である。

$$u_{ij}=\begin{cases}0 & (j\text{ が }i\text{ に対して誤答})\\ 1 & (j\text{ が }i\text{ に対して正答})\end{cases} \tag{2.2}$$

$p(u_{ij}=1|\theta_j)$ は $p_i(\theta_j)$ とも表記する。同様に、誤答確率 $1-p(u_{ij}=1|\theta_j)$ を $p(u_{ij}=0|\theta_j)$ もしくは $q_i(\theta_j)$ とする。

ここでいう 2 パラメータとは a_i と b_i のことであり、a_i を識別力値、b_i を困難度値とよぶ。定数 D は通常 1.7 とされ、それにより式 (2.1) は累積正規分布関数に近似される。θ_j は j の能力値であり、b_i と同一次元上にある。θ_j-b_i が大きいほど j の i に対する正答確率が高く、これが 0 となる場合に関数は変曲点の 0.5 をとり、そのとき正答と誤答の確率が等しいことになる。a_i は関数の傾きにあたり、これが大きいほど能力の識別力、すなわち回答結果に対する信頼性が大きい。

図 (2.1) にロジスティック関数の概形を示す。横軸が $\theta_j - b_i$ であり、縦軸が $p_i(\theta_j)$ である。

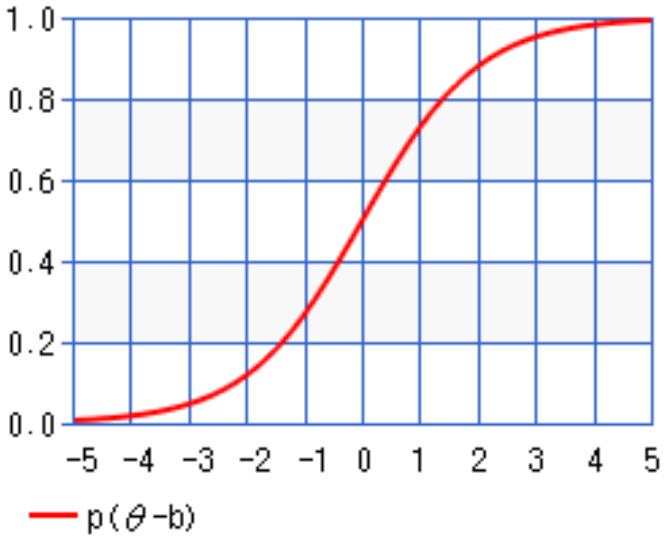

図 2.1: 2PLM におけるロジスティック関数の概形

以降、本論文では添字 j を省略する。

2.2 EAP 推定

本研究では、被検者 i の第 k 項目までの回答パターン $u_{i_1}, \ldots, u_{i_k}$（以降 u_k と表す）を得たとき、ベイズの定理より導かれる EAP（Expected a posteriori）推定の式 (2.3)[16] によって能力値の事後分布の期待値 $\hat{\theta}$ を計算する。区分求積で求めるため、離散化した形となっている。

$$p(\theta|u_k) = \frac{\sum_{h=1}^{H} X_h L(X_h|u_k) A(X_h)}{\sum_{h=1}^{H} L(X_h|u_k) A(X_h)} \tag{2.3}$$

ここで、X_h は H ある内の第 h 分点における θ の値であり、積分変数である。本研究ではガウス求積法に従いこれを決定する。$L(X_h)$ は式 (2.4) で求められる尤度

であり、回答パターンの発生確率を表す。本研究ではθの事前分布を標準正規分布と設定し、X_hにかかる重み$A(X_h)$をその確率密度関数としている。

$$L(X_h|u_k) = \prod_{l=1}^{k} p_{i_l}(X_h)^{u_{i_l}} q_{i_l}(X_h)^{1-u_{i_l}} \tag{2.4}$$

第3章　コンピュータ適応型テスト

3.1　CATのフロー

出題候補となる項目群をアイテムバンクとよぶ。被験者から得られた回答データを元に能力値 $\hat{\theta}$ を推定し、その時点で情報量の高い項目をアイテムバンクより選択して提示することを繰り返す。CAT の概念図を図 (3.1) に示す。

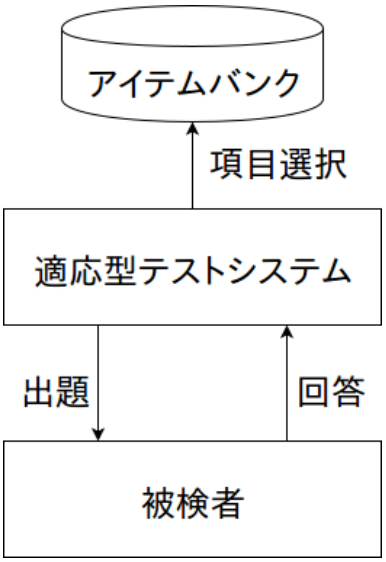

図 3.1: CAT の概念図

CAT のフローを以下に示す。本研究のシミュレーション実験もこれにしたがう。アイテムバンクに含まれる項目の数を I とする。終了条件に用いる誤差の閾値を ε とする。なお、プログラムは全て Python（バージョン 3.6.0）で記述した。

1. $k = 1$。i_k をランダムに選択する。

2. 式 (2.1) に従って回答処理を行い、その正誤によって式 (2.3) で推定値 $\hat{\theta}$ を更新する。$k = k + 1$。

3. 項目選択基準に従ってアイテムバンクより情報量が最も大きくなるような i_k を選択する。

4. $|(\hat{\theta}|u_k) - (\hat{\theta}|u_{k-1})| < \varepsilon$ か $k = I$ の場合、終了。そうでなければステップ 2 へ。

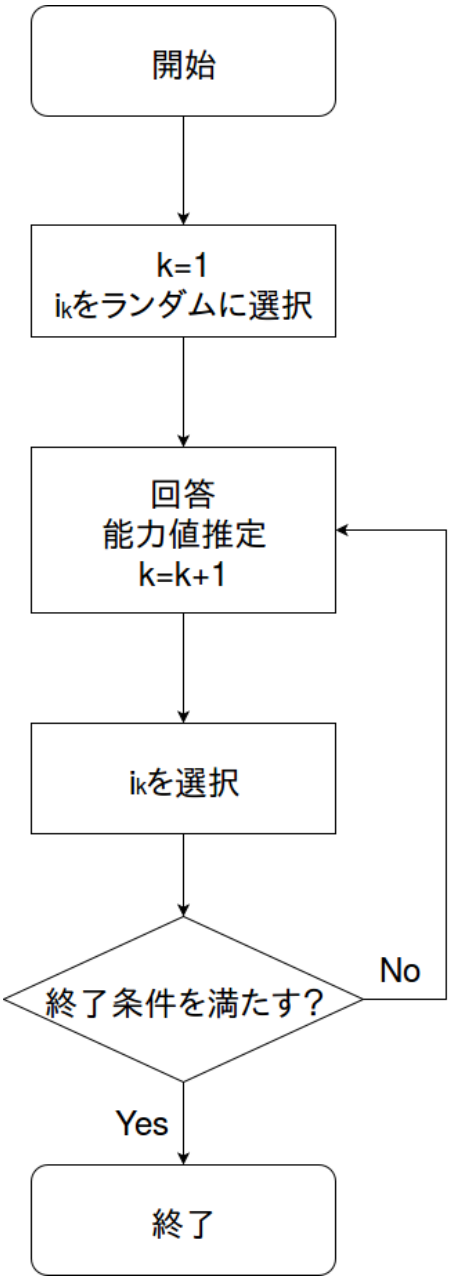

図 3.2: CAT のフローチャート

3.2 本研究で扱う情報量

第 4 章で述べる EVTI の比較対象となる情報量として、Linden[15][17] の紹介する以下の 4 つを実装した。Ueno の研究 [11] で比較されていたものと同じ情報量を採用している。

- *Fisher Information (FI)*

$$\begin{aligned} & \frac{p_i'(\hat{\theta})^2}{p_i(\hat{\theta})^2 q_i(\hat{\theta})^2} \\ = \; & D^2 a_i^2 p(\hat{\theta}_{k-1}) q(\hat{\theta}_{k-1}) \end{aligned} \tag{3.1}$$

項目反応理論の能力推定における漸近誤差は FI の逆数に一致する [18]。それゆえに測定精度を表す指標として、項目反応理論において FI が Lord & Novick(1968)[12] によって提案されており、一般的に利用される。本研究で比較対象とする他の情報量に比較して最も計算量が小さい。

- *Maximum Global-Information (MGI)*

$$\int_{\hat{\theta}_{k-1}-\delta_k}^{\hat{\theta}_{k-1}+\delta_k} K_{i_k}(\hat{\theta}_{k-1}, \theta) d\theta \tag{3.2}$$

ただし、

$$\begin{aligned} K_{i_k}(\hat{\theta}_{k-1}, \theta) &= E[\log \frac{L(\theta|u_{i_k})}{L(\hat{\theta}_{k-1}|u_{i_k})}] \\ &= p_{i_k}(\theta) \log \frac{p_{i_k}(\theta)}{p_{i_k}(\hat{\theta}_{k-1})} + q_{i_k}(\theta) \log \frac{q_{i_k}(\theta)}{q_{i_k}(\hat{\theta}_{k-1})} \end{aligned} \tag{3.3}$$

である。

MGI は Chang & Ying(1996)[13] によって提案された情報量の式である。K はカルバック・ライブラー情報量であり、2 つの分布が異なる程度を示す。本研究では δ_k を定数 0.5 とした。

- *Likelihood-Weighted Information (LWI)*

$$\int_{-\infty}^{\infty} L(\theta|u_{k-1}) I_{i_k}(\theta) d\theta \tag{3.4}$$

LWI は Veerkamp & Berger(1997)[14] によって提案された情報量の式である。I は式 (3.1) であり、それに重みとして θ の尤度をかけて積分したものである。この基準は現在の能力推定値 $\hat{\theta}_{k-1}$ に近い θ に最も大きな重みを置く。テ

ストの開始時には尤度関数はフラットであり $\hat{\theta}_{k-1}$ から離れた θ にも大きな重みがあるが、テストの終わり際になると尤度関数のピークが大きくなるために $\hat{\theta}_{k-1}$ にほとんどの重みが置かれるようになる。

- *Maximum Expected Posterior Weighted-Information (MEPWI)*

$$q(u_{i_k}|\hat{\theta}_{k-1}) \cdot \int J_{u_{k-1},u_{i_k}=0}(\theta)p(\theta|u_{k-1},u_{i_k}=0)d\theta + \\ p(u_{i_k}|\hat{\theta}_{k-1}) \cdot \int J_{u_{k-1},u_{i_k}=1}(\theta)p(\theta|u_{k-1},u_{i_k}=1)d\theta \tag{3.5}$$

ただし、

$$J_{u_{k-1}}(\theta) = -\frac{\partial}{\partial\theta^2}\ln L(\theta|u_{k-1})$$

である。MEPWI は Linden(1998)[15] によって提案された情報量の式である。θ の事後分布、すなわち尤度関数と事前分布の積を含むベイズ的な項目選択基準であり、予測される応答 u_{i_k} に対する事後分布の期待値が重みとなっている。

第4章　Expected Value of Test Information

4.1　EVTIの計算式

Ueno[11] は新しい情報量として、ベイズ決定理論の expected value of sample information（EVSI）に基づいた Expected Value of Test Information（EVTI）を提案している。回答後の θ の事後分布の分散が小さくなるような項目を選択させることが基本的な考え方である。そこで行われた比較シミュレーションにおいて、EVTI は他の情報量に比べて少ない項目数でかつ高精度な収束をもたらしている。項目 i_k に対する EVTI は式 (4.1) で計算される。

$$\begin{aligned}&\int_\theta [\ln p(\theta|u_{k-1}, u_{i_k}=0)p(u_{i_k}=0|\hat{\theta})p(\theta|u_{k-1}, u_{i_k}=0)\\&+\ln p(\theta|u_{k-1}, u_{i_k}=1)p(u_{i_k}=1|\hat{\theta})p(\theta|u_{k-1}, u_{i_k}=1)]d\theta\\&-\int_\theta \ln p(\theta|u_{k-1})p(\theta|u_{k-1})d\theta\end{aligned} \tag{4.1}$$

EVTI は FI の以下の問題点を解決する。

- 高い識別力をもつ特定の項目ばかりを選択する傾向がある
- テスト開始直後の推定誤差が大きく、誤差の大きな初期能力推定値に基づいて以後の項目選択を行ってしまう

事後分布の式 (2.3) が含まれる項を積分しているため積分の中に積分を含む形となり、計算量が特に大きくなることが EVTI の欠点であるといえる。

4.2　アルゴリズム効率化についての提案手法

EVTI を用いたテスト決定木生成のアルゴリズム効率化手法を提案する。項目選択にはなるべく高精度な予測情報量を用いることが望ましいが、EVTI をはじめとした積分中に事後分布を含む情報量は計算量が大きくなり、テストの受検者に負担を強いる可能性がある。その問題を解決するために Ueno ら [11][8] によって、正答・誤答で分岐する決定木を予め生成しておく手法が提案されている。図 4.1 に CAT おける決定木の例を示す。しかし、全体の計算量が提示項目数によって 2 のべき乗に比例して大きくなるため、項目選択に時間のかかる EVTI を項目選択基準として採用した場合は現実的な時間での生成が難しい。そこで、項目選択をオンラインで行う場合とオフラインで決定木を生成しておく場合について、動的計画法によるアルゴリズムの効率化方法を記す。

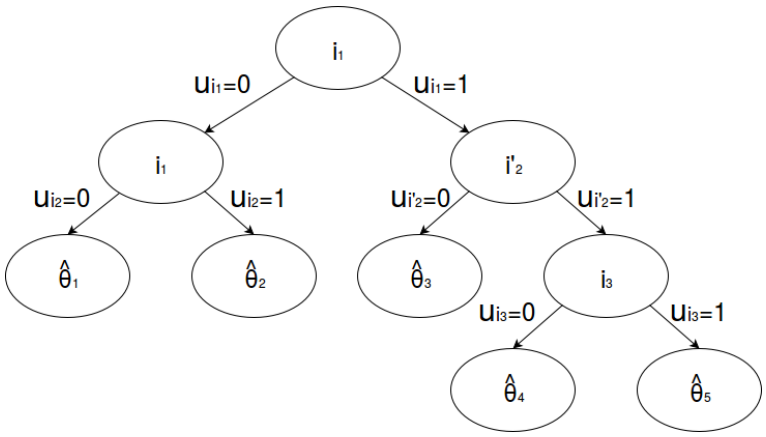

図 4.1: CAT おける決定木の例

4.2.1　オンラインで計算する場合

EVTI では現在の選択候補項目 i_k に対して誤答した場合と正答した場合の θ の事後分布を計算している。これを記憶しておくことで、i_k が決定され回答処理が行われた後にその結果によって、これのどちらかをそのまま $(\hat{\theta}|u_k)$ とすることができる。

次に、その中で計算される式 (2.4) について考える。尤度は各回答結果に対する発生確率の総乗となるので、その乗法性より式 (4.2) となる。すなわち、第 $k-1$

項目で計算された $L(X_h)$ はそのまま第 k 項目の計算で再利用できる。

$$\begin{aligned} L(X_h|u_k) &= \prod_{l=1}^{k} p_{i_l}(X_h)^{u_{i_l}} q_{i_l}(X_h)^{1-u_{i_l}} \\ &= p_{i_k}(X_h)^{u_{i_k}} q_{i_k}(X_h)^{1-u_{i_k}} \prod_{l=1}^{k-1} p_{i_l}(X_h)^{u_{i_l}} q_{i_l}(X_h)^{1-u_{i_l}} \end{aligned} \tag{4.2}$$

また、式 (4.2) は各 h について計算するため、H 個の領域を作って記憶しておく必要がある。

最後に、式 (4.1) の第 2 項は全ての選択候補項目で共通しているため、実際には省略が可能である。

4.2.2 予め木を生成しておく場合

オフラインで決定木を生成しておく場合、さらなる効率化が可能である。オンラインで計算する場合に比べ、木を生成する場合は誤答・正答時のノードが両方作られるため、$(\hat{\theta}|u_k)$ についての計算量削減の効果が大きい。深さ優先で生成する場合、最大深さを d とすると空間計算量は $O(d)$ となる。したがって、通常の動的メモリ確保による実装が可能である。

木を生成する場合における提案アルゴリズムの擬似コードを Algorithm 1 に記述する。なお、事後分布 (2.3) において、分子の計算と同時に分母を計算することができるが、簡単のためその処理の記述は省略する。

Algorithm 1 EVTI を用いた決定木構築

1: **for** X
2: 　$\hat{\theta}_k$ =mem_p
3: **if** $|\hat{\theta}_k - \hat{\theta}_{k-1}| < \varepsilon$ **then**
4: 　展開終了
5: **for** I
6: 　**for** X
7: 　　$\mathrm{mem}_L[h]$ を元に候補 i_k についての $u_{i_k} = 0, 1$ それぞれに対する $L(X_h|u_k)$ を計算し、$\mathrm{mem}_{Ltmp}[h]$ に記憶
8: 　　$\mathrm{mem}_{Ltmp}[h]$ を元に $u_{i_k} = 0, 1$ それぞれに対する $p(\theta|u_{i_k})$ を計算し、$\mathrm{mem}_{ptmp}[h]$ に記憶
9: 　mem_{ptmp} を元に EVTI を積分計算
10: 　**if** EVTI が更新された **then**
11: 　　候補 i_k を更新
12: 　　mem_L=mem_{Ltmp}
13: 　　mem_p=mem_{ptmp}
14: 決定された i_k をアイテムバンクから取り除く
15: $\hat{\theta}_k$、$k = k + 1$、$u_{i_k} = 0$ に対する mem_L と mem_p を渡して子ノードを展開
16: $\hat{\theta}_k$、$k = k + 1$、$u_{i_k} = 1$ に対する mem_L と mem_p を渡して子ノードを展開

第5章　実験設定

第 3 章で述べたフローにしたがって、CAT のシミュレーション実験を行った。決定木を生成しない通常の CAT である。同じく第 3 章で説明した 4 つの情報量と、第 4 章で説明した EVTI を比較する。出力したデータは誤差収束時の標準誤差（RMSE）および提示項目数である。

5.1　実験 1

先行研究の再現を主目的として実験 1 を行った。条件を以下に示す。実験結果として推定値と真値の誤差が小さかったとしても収束が比較的遅くなり情報量の性能比較が難しくなる、といった場合が考えられるため、終了条件に用いる閾値 ε を 0.0010 と 0.0005 の 2 通り設定する。この点は $\varepsilon = 0.0010$ のみを用いていた先行研究 [11] と異なる。

- $I = 100, 500, 1000$
- 試行回数:100
- 実験 1-1:$\varepsilon = 0.0010$
- 実験 1-2:$\varepsilon = 0.0005$

積分には区間 $[-2.0, 2.0]$ で分点数 41 のガウス求積を用いる。アイテムバンクは a を平均 0.0・標準偏差 0.4 の対数正規分布、b を平均 0.0・標準偏差 1.0 の正規分布より生成した項目からなる。被検者の真の能力値 θ は b と同様に標準正規分布より生成する。

5.2 実験2

実際の運用上、テストが繰り返されることによって、FIなどの選択基準から考えられるようにaが大きい項目ばかりが何度も曝露されアイテムバンク上から除外され、aが小さい項目にアイテムバンクが偏るといった状況がありうる。偏ったアイテムバンクから項目を選択するテストであっても安定して高精度な能力推定を行うことができるような情報量があれば有用であるといえる。

先行研究 [11] ではEVTIおよびMEPWIを用いた場合にaの大きな項目の曝露率が他に比べて大きく低かったことから、アイテムバンクに含まれる項目のaが小さい場合はそれら情報量の相対的な精度がより高いのではないかという仮説を立てる。それを検証するため、aが大きい項目を除外して項目を生成し実験2を行った。各実験の条件の詳細を以下に示す。その他の条件は実験1と同じである。

- 実験2-1:$a < 1.0, \varepsilon = 0.0050$
- 実験2-2:$a < 0.7, \varepsilon = 0.0050$

項目の識別力を落としたことによって、予備実験にて予想以上に収束の遅れが観察されたため、閾値εを0.0050に上げている。実際に実施するテスティングにおいては当実験結果よりもさらに提示項目数が少なくなるが、結果のばらつきを小さくするために本研究のシミュレーション実験では収束条件となる閾値を小さくして収束時の誤差を記録する。

5.3 実験3

実験2と同様に、アイテムバンクに含まれる項目にbが真のθの近傍のものが含まれない場合、推定精度がどのように変化するのかを確認したい。本研究では$b < \theta - 0.3, \theta + 0.3 < b$の場合について検証する。同一アイテムバンクで試行を繰り返す都合上で真のθを3つの値に固定して実験3を行った。その他の条件は実験1と同じである。

- 共通条件:$b < \theta - 0.3, \theta + 0.3 < b$
- 実験 3-1:$\theta = 0.0$
- 実験 3-2:$\theta = 0.5$
- 実験 3-3:$\theta = 1.0$

第6章　実験結果と考察

6.1　実験1

表6.1,6.2に示す実験1のデータを観察する。各条件でRMSEが最も小さかった情報量を挙げる。$\varepsilon = 0.0010$のとき、

- $I = 100$:EVTI
- $I = 500$:LWIとEVTI
- $I = 1000$:MGIとEVTI

であった。また$\varepsilon = 0.0005$のとき、

- $I = 100$:EVTI
- $I = 500$:EVTI
- $I = 1000$:MEPWIとEVTI

であった。提示項目数が最も小さかった情報量を挙げる。

- すべての条件:EVTI

であった。

表 6.1: 実験結果 1-1

ε=0.0010		
Method	RMSE	Number of items (SD)
I=100		
FI	0.20	81.16 (15.12)
MGI	0.15	82.12 (11.62)
LWI	0.17	82.00 (14.98)
MEPWI	0.15	72.58 (12.20)
EVTI	0.13	69.44 (13.82)
I=500		
FI	0.12	173.22 (32.79)
MGI	0.14	138.35 (40.80)
LWI	0.11	180.11 (36.46)
MEPWI	0.12	135.26 (35.58)
EVTI	0.11	132.26 (33.36)
I=1000		
FI	0.15	176.51 (48.25)
MGI	0.10	135.80 (44.56)
LWI	0.12	174.40 (53.09)
MEPWI	0.12	139.93 (43.31)
EVTI	0.10	127.01 (40.57)

表 6.2: 実験結果 1-2

ε=0.0005		
Method	RMSE	Number of items (SD)
I=100		
FI	0.19	86.85 (14.42)
MGI	0.18	88.45 (9.58)
LWI	0.19	89.56 (9.33)
MEPWI	0.18	85.01 (22.05)
EVTI	0.17	82.31 (14.33)
I=500		
FI	0.13	261.73 (49.56)
MGI	0.19	199.93 (56.00)
LWI	0.12	255.98 (51.06)
MEPWI	0.12	193.70 (55.29)
EVTI	0.11	172.59 (74.77)
I=1000		
FI	0.11	309.67 (102.90)
MGI	0.11	209.71 (67.32)
LWI	0.12	281.69 (104.70)
MEPWI	0.10	216.55 (100.36)
EVTI	0.10	204.21 (95.33)

実験 1 では、特別な条件を与えずに各情報量における $\hat{\theta}$ の推定精度および収束速度を比較した。実装方法と試行回数が異なるためか先行研究ほどの大きな差は見られなかったが、EVTI と MEPWI が他の情報量と比較して誤差が小さく、提示項目数も小さく出ることがわかった。これらの情報量は事後分布の積分を含むことから、その計算量の大きさと引き換えにより適切な項目提示ができる傾向にあると考えられる。それら 2 つを比較すると MEPWI よりも EVTI のほうが誤差が僅かに小さく、EVTI の有効性を確認することができた。$\varepsilon = 0.0010$ と $\varepsilon = 0.0005$ の条件ともに、$I = 500$ と $I = 1000$ を比較すると提示項目数の見られず、一方で $I = 1000$ のほうが誤差が小さくなっている。これによってアイテムバンクが大きくなるとある程度で収束速度が打ち止めになり、そこからは誤差が小さくなっていくという振る舞いが観察された。今回比較した 5 つの情報量の中で、MGI は推定値の変動幅が小さくなるタイミングが早いためか、提示項目数が少ない代わりに誤差が不安定に出ることが確認できた。MGI を用いたテストを行う場合は ε をさらに小さくするなどの工夫を施すことが効果的である可能性がある。

6.2 実験 2

表 6.3,6.4 に示す実験 2 のデータを観察する。各条件で RMSE が最も小さかった情報量を挙げる。$a < 1.0$ のとき、

- $I = 100$:MEPWI
- $I = 500$:MEPWI と EVTI
- $I = 1000$:MGI と EVTI

であった。$a < 0.7$ のとき、

- $I = 100$:MEPWI と EVTI
- $I = 500$:MEPWI と EVTI
- $I = 1000$:EVTI

であった。提示項目数が最も小さかった情報量を挙げる。

- $a < 1.0$ の $I = 100, 500$ と $a < 0.7$ の $I = 500$:MEPWI
- その他の条件:EVTI

であった。

表 6.3: 実験結果 2-1

$a < 1.0$		
ε=0.0050		
Method	RMSE	Number of items (SD)
I=100		
FI	0.25	70.02 (12.33)
MGI	0.23	67.76 (14.57)
LWI	0.24	69.46 (12.94)
MEPWI	0.21	62.20 (11.37)
EVTI	0.22	63.51 (11.49)
I=500		
FI	0.16	119.02 (23.14)
MGI	0.15	116.13 (20.64)
LWI	0.15	119.98 (20.90)
MEPWI	0.14	108.19 (18.99)
EVTI	0.14	110.62 (19.27)
I=1000		
FI	0.18	134.08 (28.91)
MGI	0.13	132.92 (18.00)
LWI	0.15	134.26 (27.40)
MEPWI	0.14	115.27 (27.03)
EVTI	0.13	114.93 (26.88)

表 6.4: 実験結果 2-2

$a < 0.7$		
ε=0.0050		
Method	RMSE	Number of items (SD)
I=100		
FI	0.27	89.28 (11.23)
MGI	0.27	86.82 (14.93)
LWI	0.25	87.28 (14.91)
MEPWI	0.23	77.01 (11.38)
EVTI	0.23	76.65 (10.92)
I=500		
FI	0.20	172.27 (32.84)
MGI	0.20	164.44 (40.13)
LWI	0.20	169.68 (37.09)
MEPWI	0.18	148.96 (31.53)
EVTI	0.18	151.07 (30.50)
I=1000		
FI	0.21	196.97 (46.26)
MGI	0.20	190.94 (44.51)
LWI	0.18	197.16 (43.01)
MEPWI	0.18	161.14 (40.56)
EVTI	0.16	153.48 (39.38)

実験2では、εを大きくした上でアイテムバンク全体の項目の識別力aを落として各情報量を比較した。アイテムバンク生成の条件に$a < 1.0$と$a < 0.7$の2段階を設けたが、実験1の結果と比較して期待していたような情報量間の誤差の開きは見られず、今回の実験ではaの上限を小さくするにつれてすべてが等しく増加しているといえる。識別力の大きな項目のない中でEVTIが特別なパフォーマンスを発揮するといったような結果は得られなかった。

6.3 実験3

表6.5,6.6,6.7に示す実験3のデータを観察する。各条件でRMSEが最も小さかった情報量を挙げる。$\theta = 0.0$のとき、

- $I = 100$:EVTI
- $I = 500$:MEPWIとEVTI
- $I = 1000$:LWIとMEPWIとEVTI

であった。$\theta = 0.5$のとき、

- $I = 100$:MGIとLWIとEVTI
- $I = 500$:LWIとMEPWIとEVTI
- $I = 1000$:LWIとEVTI

であった。$\theta = 1.0$のとき、

- $I = 100$:MEPWIとEVTI
- $I = 500$:EVTI
- $I = 1000$:FIとMEPWIとEVTI

であった。提示項目数が最も小さかった情報量は、

- $\theta = 0.0$ の $I = 500, I = 1000$ と $\theta = 1.0$ の $I = 100, 500, 1000$:MGI
- $\theta = 0.5$ の $I = 500$:MEPWI
- その他の条件:EVTI

であった。

表 6.5: 実験結果 3-1

$\theta = 0.0, b < -0.3, 0.3 < b$		
ε=0.0010		
Method	RMSE	Number of items (SD)
I=100		
FI	0.14	87.84 (3.02)
MGI	0.15	82.70 (6.12)
LWI	0.14	87.88 (3.20)
MEPWI	0.15	80.49 (2.98)
EVTI	0.13	79.22 (2.36)
I=500		
FI	0.08	207.48 (19.59)
MGI	0.10	166.03 (21.16)
LWI	0.08	207.00 (19.17)
MEPWI	0.07	172.39 (18.55)
EVTI	0.07	167.51 (17.21)
I=1000		
FI	0.08	204.74 (27.75)
MGI	0.08	149.89 (27.93)
LWI	0.07	203.92 (26.01)
MEPWI	0.07	169.02 (25.33)
EVTI	0.07	168.82 (26.10)

表 6.6: 実験結果 3-2

$\theta = 0.5, b < 0.2, 0.8 < b$		
ε=0.0010		
Method	RMSE	Number of items (SD)
I=100		
FI	0.17	91.51 (2.08)
MGI	0.15	87.01 (4.03)
LWI	0.15	91.31 (1.59)
MEPWI	0.14	90.33 (2.05)
EVTI	0.15	86.08 (2.37)
I=500		
FI	0.10	191.49 (29.10)
MGI	0.10	152.15 (28.69)
LWI	0.07	195.47 (30.06)
MEPWI	0.07	142.32 (28.41)
EVTI	0.07	144.99 (27.05)
I=1000		
FI	0.07	204.60 (31.99)
MGI	0.08	166.54 (36.94)
LWI	0.06	198.71 (27.59)
MEPWI	0.07	141.39 (29.20)
EVTI	0.06	143.71 (28.86)

表 6.7: 実験結果 3-3

$\theta = 1.0, b < 0.7, 1.3 < b$		
ε=0.0010		
Method	RMSE	Number of items (SD)
I=100		
FI	0.16	80.85 (1.31)
MGI	0.17	78.37 (2.39)
LWI	0.16	80.59 (1.50)
MEPWI	0.15	79.11 (1.79)
EVTI	0.15	80.07 (1.56)
I=500		
FI	0.09	189.93 (15.21)
MGI	0.09	133.25 (20.20)
LWI	0.09	191.28 (14.72)
MEPWI	0.08	162.55 (16.32)
EVTI	0.07	152.10 (15.95)
I=1000		
FI	0.07	206.40 (26.93)
MGI	0.08	110.32 (26.63)
LWI	0.08	201.31 (28.04)
MEPWI	0.07	163.48 (26.11)
EVTI	0.07	160.01 (25.58)

実験3では、シミュレーションを通しての真のθを固定し、θの近傍の困難度bをもつ項目がないようなアイテムバンクを生成し、3通りの比較を行った。これによって情報量間の差異がより大きくなる可能性が予想されたが、実験2の結果と同様に今回の実験結果からも特にそのような性質を見出すことができなかった。同じεを設定した実験1の結果と比較すると、ある程度の回答を繰り返した後であっても推定値$\hat{\theta}$に対して適切なbをもつ項目が乏しくなるために$\hat{\theta}$はなかなか収束しないが、収束時点での誤差は小さく出ると見られる。いずれの情報量を用いても誤差が同程度であったことから、EVTIなど複雑な情報量に関してはbに強く依存している、すなわちbが偏っていないアイテムバンクを必要とする可能性があると考えられる。

以上のように、アイテムバンクの項目パラメータが偏ることによって情報量ごとのパフォーマンスに差が生まれるかどうかを検証する実験を行ったが、全体を通してそのような傾向は発見されなかった。

第7章　おわりに

7.1　結論

CAT において、テスト実施時の項目選択時間を短縮するためにオフラインで項目決定木を生成する手法 [8] と、高精度な予測情報量 EVTI[11] が提案されている。しかし、EVTI の計算時間が非常に大きい。推定精度と計算時間がトレードオフの関係にあるため、オフラインで木を生成する場合でも十分な精度のテストを構成するためには膨大な時間がかかるという問題があった。

そこで本研究では、EVTI を用いて決定木を生成する場合のアルゴリズム効率化手法を第 4 章で提案した。これを適用することによって、現実的な時間で十分に大きな決定木を生成することができると考えられる。

事前に決定木の形でテストを構成することによって被検者に負担をかけることなく精度を高めることができる。つまり、提示項目数を減らすことができる。その結果として以下のメリットが得られる。

- テスト時間が短縮されることによって被検者の負担が軽減される
- 項目の曝露率が下がることによってアイテムバンクの項目を長期間用いることができる

次に、EVTI と他の 4 つの情報量を CAT の項目選択基準として実装し、それらが活かされる条件を発見すべく、アイテムバンクに以下の条件を与えて CAT シミュレーション実験を行った。

- 通常のアイテムバンク
- a の小さな項目のみを含むアイテムバンク
- 被検者の θ から外れた b をもつ項目のみを含むアイテムバンク

このように提示候補となる項目群のパラメータに偏りをもたせてシミュレーションを繰り返した結果、以下の結果が確認された。

- 全体として概ねEVTIの誤差と提示項目数が最も小さくなった
- 同様に計算量の大きなMEPWIの誤差と提示項目数が次いで小さくなった
- 条件を付加したアイテムバンクを用いた場合に情報量間のパフォーマンスに大きな差は見られなかった

7.2 今後の課題

今後の課題として、EVTIをはじめとした情報量のさらなる詳細な特性を明らかにすることができれば、それらを組み合わせたハイブリッド的な情報量を新たに提案し、より推定精度を高いものにすることができると考えられる。また、今後の研究によってEVTIが通常よりも有効となる条件が明らかになった場合、その条件に適合する状況下において項目選択基準としてEVTIを採用した効率化CATアルゴリズムを積極的に用いることによって、テスティングパフォーマンスの向上に大きく寄与すると考えられる。

謝辞

本研究を行うにあたり、親切にご指導を頂いた植野真臣教授に厚く御礼を申し上げます。ゼミ発表中のコメントが大変励みになりました。並びに川野秀一准教授にはシミュレーション手法などに関わる提言を頂き、感謝しております。また、プログラミングまわりの相談に際して宇都雅輝助教と東京学芸大学の宮澤芳光助教には重要な助言を賜りました。御陰様で問題点を修正することができました。皆様ありがとうございます。最後に、困難な学生生活を最後まで支えて頂いたすべての家族・友人に謝意を表します。

参考文献

[1] 植野真臣, 永岡慶三 (2009),e テスティング, 培風館.

[2] K. Kato (2016), “Measurement Issues in Large-Scale Educational Assessment”, The Annual Report of Educational Psychology in Japan 55, pp.148-164.

[3] 独立行政法人情報処理推進機構, “IT パスポート試験”, https://www3.jitec.ipa.go.jp/JitesCbt/.

[4] 公益社団法人医療系大学間共用試験実施評価機構, “臨床実習開始前の「共用試験」第 13 版 (平成 27 年度)”, http://www.cato.umin.jp/e-book/13/index.html.

[5] 谷澤明紀, 本多康弘 (2014), “情報処理技術者試験における e テスティング”, 日本テスト学会第 12 回大会発表論文抄録,33(2), pp.5457.

[6] 仁田善雄, 齋藤宣彦, 後藤英司, 高木康, 石田達樹, 江藤一洋 (2014), “医療系大学間共用試験における e テスティング”, 日本テスト学会第 12 回大会発表論文抄録集, pp.5859.

[7] 池田 央, 柳井晴夫, 藤田恵璽, 繁桝算男 (1992), 教育測定学〈下巻〉, 学習評価研究所.

[8] M. Ueno, P. Songmuang (2010), “Computerized Adaptive Testing based on Decision Tree”, The 10th IEEE International Conference on Advanced Learning Technologies, pp.191-193.

[9] D. Delgado-Gomez, E. Baca-Garcia, D. Aguado, P. Courtet, J. Lopez-Castroman (2016), “Computerized Adaptive Test vs. decision trees: Development of a support decision system to identify suicidal behavior”, Journal of Affective Disorders 206 pp.204-209.

[10] D. Delgado-Gomez, Juan C. Laria, Diego Ruiz-Hernandez (2019), "Computerized adaptive test and decision trees: A unifying approach", Expert Systems With Applications 117 pp.358-366.

[11] M. Ueno (2013), "Adaptive Testing Based on Bayesian Decision Theory", Artificial Intelligence in Education 2013, pp.712-716.

[12] F.M. Lord & M.R. Novick (1968), Statistical Theories of Mental Test Scores, Addison-Wesley Educational Publishers Inc.

[13] H. Chang, Z. Ying (1996), "A global information approach to computerized adaptive testing", Applied Psychological Measurement 20(3), pp.213-229.

[14] W.J.J. Veerkamp, M.P.F. Berger (1997), "Some New Item Selection Criteria for Adaptive Testing", Journal of Educational and Behavioral Statistics 2(2), p.p.203-226.

[15] W.J. van der Linden (1998), "Bayesian Item Selection Criteria for Adaptive Testing", Psychometrika 63(2), pp.201-216.

[16] 豊田秀樹 (2005), 項目反応理論 [理論編], 朝倉書店.

[17] W.J. van der Linden (2000), Computerized Adaptive Testing Theory and Practice, Kluwer Academic Publishers.

[18] F.M. Lord (1980), Applications of Item Response Theory to Practical Testing Problems, Routledge.