JSiSE Research Report
vol.38,no.2(2023-7)

学習履歴と教員ポリシー双方を考慮した CBT の難易度分類アルゴリズムの構築

荒関虹希*1，上野春毅*2，小松川浩*1
*1 公立千歳科学技術大学大学院 理工学研究科　*2 公立千歳科学技術大学 理工学部

Construction of a CBT Difficulty Classification Algorithm Considering Both Learning History and Teacher Policy

Koki Araseki *1，Haruki Ueno*2，Hiroshi Komatsugawa*1
*1 Graduate School of Science and Engineering，Chitose Institute of Science And Technology
*2 Faculty of Science and Technology, Chitose Institute of Science and Technology

We construct and evaluate a difficulty classification algorithm for a wide variety of exercises and test materials consisting of images and texts managed by each material provider. In the previous study, classification was performed by using test theory to supplement missing values based on learning history. In the present study, we use machine learning, including learner's test results with missing values and the learning policy designated by teachers.

キーワード: 個別最適化，テスト理論，機械学習，難易度分類，教員ポリシー

1. はじめに

近年，初等中等教育機関での Giga スクール構想を通じて，学校と家庭を繋ぐ中での個別最適な学習を図ることが求められている．そのための学習支援システムとして，オンライン上で演習やテスト問題を管理し，個々の学習者に適した教材提供する Computer-based Testing(CBT)の活用が期待されている．その一例として，文部科学省が提供する CBT システム（MEXCBT）や，商用の AI ドリルが挙げられる．しかし，MEXCBT(1)は出版社や専門家が提供する複数の教材提供が提供した体系的な無償の問題群を，管理者が一元的に管理して提供するものであり，適応的に問題を出題する Computer-Adaptive-Testing(CAT)形式にはなっていないため，個別最適な学習支援の観点で課題が残る．一方，商用の AI ドリルには，CAT 機能は実装されているが，提供教材がベンダー固有の有償教材となっている．そのため，各地域特性や学校の置かれている状況が異なる現在の日本の教育事情を考慮すると，必ずしも現場教員の日頃の教育活動に即した個別最適な学習支援を図れているとはいえない．

現場教育に即した個別最適な学習支援の観点では，現場に近い多くの教材提供者（都道府県・市町村単位の教育委員会や複数の学校コミュニティ）が柔軟に無償の CAT 向けの教材を提供できることが望ましい．しかし，教材を更新する際に，都度各演習やテストの難易度の設定が必要となり，現場教員の負担という観点で課題が残る．さらに，信頼性・妥当性の高い真正な評価を意識したテスト理論に基づく（公開性が担保されている）難易度の設定を各教育現場レベルで行うことは，現実的ではない．

そこで本研究では，テスト理論と機械学習アルゴリズムを活用して，教育現場が日頃から保有する多種多様なテストや演習問題群に対する難易度自動分類アルゴリズムの構築を目的とする．それにより，教育委員会や学校といった各教材提供者内で運用できる個別最適な学習向けの CAT システムの実現を目指していく．

実証的な研究を進めるため，現場教員が日頃活用している実際の CAT における問題の難易度の自動調整を想定し，北海道内に 9 万人規模のユーザを持つ CIST-Solomon(2)の CAT を基盤システムとして扱う．当該システムにおけるテストは，テスト理論による CAT のロジック

で稼働している.

2. 先行研究と本研究の位置づけ

2.1 テスト理論

難易度分類に関する先行研究では，テスト理論が採用されることが多い．テスト理論とは，テストに関する知識体系や能力判定のための統計モデルの理論であり，テストの標準化・尺度化を主な目的としている．テスト理論は，項目の難易度・識別力を測定する項目反応理論（Item Response Theory ：IRT）[3]や潜在特性を順序尺度にした潜在ランク理論（Latent Rank Theory ：LRT）[4]が代表的である.

2.2 先行研究の例

金西らの研究[5]では，初年次教育の理系基礎科目を対象とした大規模な演習教材群に対し,CAT を用いた個別最適な学習システムの仕様検討を行った．その際，学習システムにおける問題の作成と解答履歴の収集を行い，IRT を用いた問題の難易度分類を試みた．その結果，IRT を用いて正確な難易度分類を行うためには，様々な理解状況の学習者から学習履歴を収集する必要があるとした.

阿部の研究では，CIST-Solomon の回答履歴と Exametrika[6](LRT を実装可能なシステム）を用いて推定された IRP をもとに難易度更新を行った．なお難易度はレベル 1～7 の 7 段階としていた．具体的な阿部のアルゴリズムを図 1 に示す．最初に CIST-Solomon のデータベースからある単元の回答履歴を抽出し，Exametrika を用いて RMP（受験者のランク確率）を推定する．このとき，抽出した回答履歴には欠損値があるため，阿部が作成したシステムでは RMP をもとに正答する確率を仮定し，重み付けを行ったうえで欠損値を補完していた．その後，補完済みの回答履歴（Excel ファイル）を再度 Exametrika で分析し，推定された IRP を各問題の更新された難易度としていた.

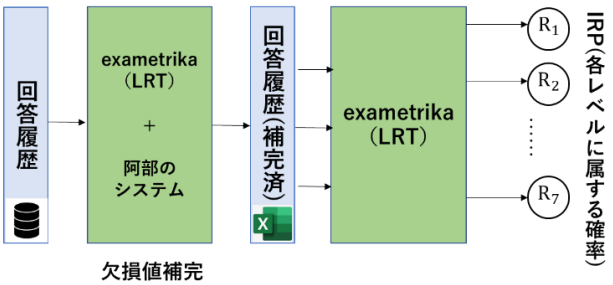

図 1 阿部のアルゴリズム

2.3 本研究の位置づけ

テスト理論では，真正な評価を意識すると各演習問題に大規模なテスト履歴に基づき複数のパラメータ推定が必要となる．一方，近年初等教育での Giga スクール推進や高等教育でのポストコロナでのオンライン教育の普及に伴い，教育機関独自に e ラーニングの教材運用する事例が増えている．こうした学習コミュニティでは，現場教員が適宜教材管理することから，演習問題の学習履歴も均質とならない．そこで，本研究では教材提供コミュニティ単位での教育事情（内容）に応じた個別最適な学習の提供を可能にするために，テスト理論に基づく CAT に対し，教員の教育ポリシーを反映した機械学習アルゴリズムで置き換えたロジック構築を図り，その有用性を評価する.

本研究では，先行研究が持つ課題点を解決するために，汎用的な実稼働システムである CIST-Solomon の学習履歴を事例として扱い，ルーブリックを入力パラメータとして追加した全結合型ニューラルネットワーク(Neural Network ：NN)のアルゴリズム構築を目指す．このとき，LRT が IRT よりも分類において優れた性能を表すため，阿部の先行研究をもとにアルゴリズムの構築を図る.

3. アルゴリズムの構築と評価

3.1 難易度の定義

本研究では，CIST-Solomon における理解度テスト(CAT)の学習履歴をもとに，難易度分類モデルの作成を行う．そのため，難易度更新の対象として，本学の講義内で理解度テストを例年実施している「アルゴリズムとプログラミング」から，「再帰処理」と「ソートアルゴリズム」の 2 単元を選択した．ただし，この 2 単元は先行研究である阿部のアルゴリズムによって難易度が更新された単元である．そのため，この 2 単元の難易度は，問題を登録した専門家の意見や LRT による推定等，多角的観点から定められたものである．また，2 単元を対象とした理由は，検証の際に単元による相違点を比較するためである.

3.2 難易度分類アルゴリズムの構築

阿部の先行研究では，CIST-Solomon における理解度テストの問題が真正な評価に基づくと仮定し,LRT を活用することで潜在ランク R1～R7 を求め，最大となる潜在ラン

クを問題の難易度として推定していた．一方で，学生の自習で活用されるシステムでは，学習者毎に回答する問題も異なる．また，適宜教員による問題の追加も行われることを考慮すると，未解答の回答履歴（欠損値）も多い．これに対して，阿部の研究では一般的なNNで用いられる最尤推定による，確率に基づいたデータ補完が行われていた．この場合，LRTの構造自体が教師なしNNであるNeural Network SOM（SOM）[7]と同じであるため，上記のデータ補完を含めて，データ補完を含めた全結合型のNNを用いた機械学習モデリングで表現できる可能性がある．一方で，実稼働システムでは，真正なデータが揃わないデータに対しては，表1に示す教員の教育ポリシー（ルーブリック）に沿って問題の難易度を決めている．そこで，本研究ではNNの入力パラメータとしてルーブリックを追加し，教師データを教員により更新された難易度とすることで，全結合型NNによる難易度分類アルゴリズムの構築を図ることとした．

表 1　レベルとルーブリックの対応表

レベル	ルーブリック
1	初級
2	
3	中級
4	
5	
6	上級
7	

第一に，図2に示すような阿部のアルゴリズムにNNを組み合わせたアルゴリズムを構築した（以下，アルゴリズム1とする）．すなわち，アルゴリズム1ではLRTをNNの入力層として明示的に利用した．さらにルーブリック情報を入力層に加え，3層構造のNNを構築することで，教員の意見が一定程度反映される難易度更新アルゴリズムとした．

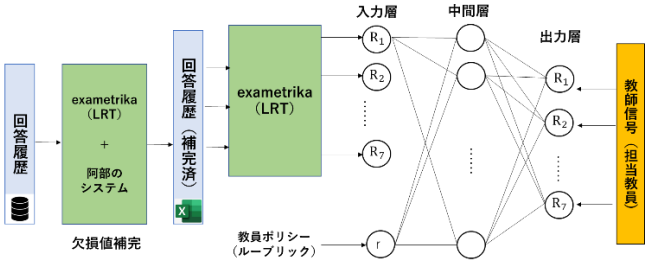

図 2　アルゴリズム1

第二に，図3に示すようなExametrikaと阿部のシステムによる欠損値補完とNNを組み合わせたアルゴリズムを構築した（以下，アルゴリズム2とする）．アルゴリズム1では，補完後のデータに対してExametrikaを通すことでLRTによる潜在ランクを推定しNNの入力に利用した．一方で，アルゴリズム2では，補完後のデータから算出された基本統計量をNNの入力に用いることで，テスト理論に依存しないアルゴリズムを目指した．

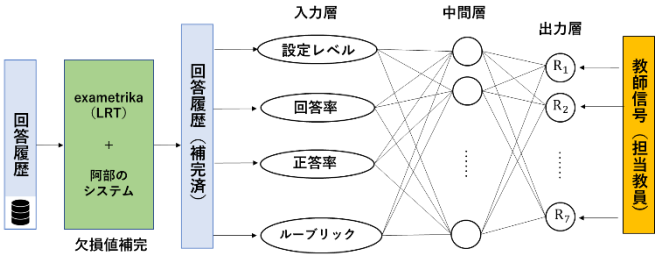

図 3　アルゴリズム2

第三に，図4に示すような学習履歴の欠損値補完を行わず，NNのみによるアルゴリズムを構築した（以下，アルゴリズム3とする）．アルゴリズム2では欠損値補完を行ったうえでNNを適用していたが，アルゴリズム3では，Exametrikaによる補完を行わず欠損値が存在するデータから算出された基本統計量をNNに用いた．そのため，LRTに依存しないアルゴリズムとなり，将来的にシステムのみによる自動更新を可能にした．

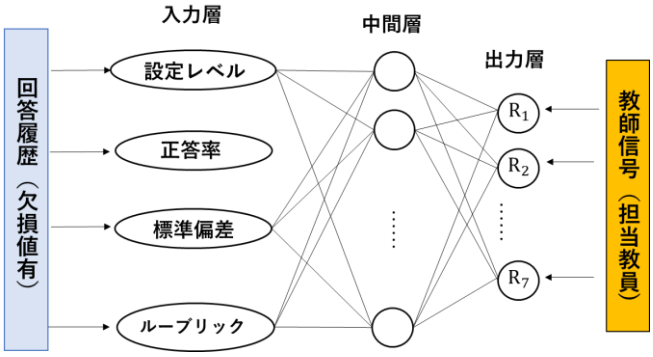

図 4　アルゴリズム3

3.3　モデルの作成

前述の阿部のアルゴリズムとアルゴリズム1～3を用いて，「再帰処理」と「ソートアルゴリズム」の2単元において計8種類のモデルを作成した．本研究におけるモデルの教師信号とは，阿部のアルゴリズムによって2021年の解答履歴をもとに分類された問題を，担当教員が単

元ごとに相対評価することで再定義した難易度である．そのため，学習時における入力データは2021年の解答履歴とすることが望ましい．しかし，2021年の解答履歴のみを用いた場合，「再帰処理」の問題数は79件，「ソートアルゴリズム」の問題数は93件のみであり入力データが不足する．そのため，本研究では学習時の入力データを2017～2020年の解答履歴，教師信号を教員により更新された難易度とすることでデータセットを4倍にかさ増しした．また，推論時の入力データは2021年の学習履歴とし，教師信号は学習時と同じく教員により更新された難易度とした．この際，2017 ～2020 年の各単元の問題は2021年のものと同質であり，担当教員は主に問題の内容とルーブリックをもとに難易度を更新しているため，2017～2020 年においても担当教員は同様の更新を行うと仮定した．NN のハイパーパラメータに関しては，「再帰処理」のモデル作成時に調整を行い，epoch数を1000，バッチサイズを1として固定した．

3.4 モデルの評価

更新された難易度を評価する際に，観点により難易度の解釈が異なるため，一般的に「正しい難易度」とする指標は存在しない．そこで本研究では，「アルゴリズムとプログラミング」の担当教員に難易度更新における協力を依頼し，その難易度を一時的に「正しい難易度」と定義した．そのため評価における精度とは，開発したモデルにより推定された難易度と担当教員が更新した難易度の一致度を示している．

表2に各モデルによる単元ごとの評価を示す．ここでレベル毎の問題数の分布を測るために，標準偏差も求めた．精度に着目すると，先行研究である阿部のアルゴリズムではどちらの単元も約 6～7 割の精度であったが，本研究で提案したアルゴリズム1～3のうち，「再帰処理」ではアルゴリズム1･3が約8割，「ソートアルゴリズム」ではアルゴリズム2･3が約7～8割と高い精度を示した．特に，潜在ランクを一切扱わないアルゴリズム3は，2つの単元において最も高い精度を示していた．

標準偏差に着目すると，「再帰処理」ではアルゴリズム1が最も小さく，「ソートアルゴリズム」ではアルゴリズム3が最も小さかった．しかし，実際に分類された問題を分析すると，問題数が0～2問ほどしかしないレベルも存在した．そのため，理解度テストを構成するうえでどのアルゴリズムも適切な更新・分類が行われたとはいえない．

表 2　モデルによる単元ごとの評価

	再帰処理		ソートアルゴリズム	
アルゴリズム	精度(%)	標準偏差	精度(%)	標準偏差
阿部のアルゴリズム	65.8	7.12	62.4	14.1
アルゴリズム1	79.8	6.39	54.8	22.6
アルゴリズム2	58.2	7.56	73.1	16.6
アルゴリズム3	84.8	7.68	78.5	10.2

4. 考察と今後の展望

以上より，本研究における精度の観点からは，アルゴリズム3すなわちNNのみで構成されたアルゴリズムが最も優れていたと推測される．これは潜在ランク理論が大規模かつ欠損のないデータを前提に提唱されているものであり，本学における CAT 形式の理解度テストではデータ欠損やデータ不足が見られるため，充分に作用しなかったためと考えられる．しかし，本研究では入力データとして，阿部の先行研究によって専門家の意見や LRT などの多角的観点から難易度分類が行われた問題を使用している．そのため，NN のみのアルゴリズムに関しても，入力データを通して LRT が作用している可能性も捨てきれない．

一方で，各レベルにおける問題数の分布の観点からは，担当教員が更新したものと比較すると，レベル間における分布の偏りがあり，なかには1問も属していないレベルが存在するなど，実稼働システムとしての課題が残る．この課題の要因としてモデリングの手法が挙げられる．改善案としては，2021年の解答履歴と担当教員により与えられた教師データのデータセットを増やすことが挙げられる．そのため，本研究では2単元に対してそれぞれモデルを作成したが，教員による難易度分類を行った対象単元を増やし，単元を越えて各レベルの分布が均等になるようにデータを混ぜて学習することで課題解決が見込まれる．また，別の要因としてルーブリック，すなわち教員ポリシーの学習効果が弱いことが挙げられる．実際に教員が問題の難易度を設定する際には，回答履歴ではなく各問題文の文脈に着目している．そのため，難易度分類をする際に文脈の類似度も考慮する必要があると考えられる．

今後の展望として，BERT を用いて問題文のベクトル表

現を算出し，NN の入力層に組み入れることで難易度分類にどのような影響があるのかを分析していきたい.

参 考 文 献

(1) 文部科学省.“文部科学省 CBT システム (MEXCBT: メクビット) について” . https://www.mext.go.jp/a_menu/shotou/zyouhou/mext_00001.html (2023 年 6 月 10 日確認)

(2) 公立千歳科学技術大学.“e ラーニングシステム CIST-Solomon ” . 公立千歳科学技術大学. http://solomon.mc.chitose.ac.jp/CIST-Shiva/Index (2023 年 6 月 10 日確認)

(3) 宇佐美 慧, 荘島 荘二郎, 光永 悠彦, 登藤 直弥. 項目反応理論 (IRT) の考え方と実践――測定の質の高いテストや尺度を作成するための技術――. 日本教育心理学会第 60 回総会発表論文集. 2018, 60, 24-25

(4) 荘島 宏二郎. ニューラルテスト理論－資格試験のためのテスト標準化理論－. 電子情報通信学会誌. 2009, 92, 1013-1016

(5) 金西 計英, 石田 基広, 戸川 聡, 高橋 暁子. 初年次教育の理系基礎科目を対象にした適応型システムの検討. 日本教育工学会論文誌. 2021, 45, Suppl, 189-192

(6) 荘島 宏二郎. ”exametrika” . SHOJIMA Kojiro’s Website . http://shojima.starfree.jp/exmk/index.html (2023 年 6 月 12 日確認)

(7) 徳高 平蔵, 藤村 喜久郎. 自己組織化マップとその応用事例. 日本神経回路学会誌. 2003, Vol.13, No.3, 147-157