

情報技術学習支援システムの開発

—学習評価のための問題文の分類に関する検討—

宮川 裕介[†] 泉 隆[†]

日本大学[†]

1. はじめに

ICT 技術の発展により、インターネットを用いて学習を行う e-Learning システムが教育機関や企業で利用されている。このシステムの利点として、時間や場所の制約がない、学習や教育の効率化を図れるといった点が挙げられる。しかし、その一方でシステムの利用頻度は利用者の学習意欲に依存するという欠点がある。学習意欲の低下は e-Learning システムが掲げる目標の達成度に大きな影響を及ぼすため、学習意欲を維持・向上させるようなシステムの構築が求められる。

学習意欲が低下する要因として、利用者個人の持つ特性を考慮していない点が挙げられる。この特性とは、学習内容に対する理解度や習熟度、学習に要する時間などである。これらの特性を考慮し、個人に合わせたテストや学習計画を提案することで利用者の学習意欲向上につながることを考える。本研究では、基本情報技術者試験^[1]を対象とした e-Learning システムから得られるデータをもとにテスト内容や学習のペースを個人化する。そして、その結果から学習評価を行い、各利用者にフィードバックすることで学習意欲の向上を目指す。

本報告では、問題の種類別および分野別分類の精度向上を目的として、形態素を用いた分類手法について検討したので述べる。

2. 適応型テストの概要と問題点

各利用者の能力に合わせて逐次的に問題を出題するものを適応型テストという。この適応型テストには、項目反応理論(Item Response Theory; IRT)やネットワーク型モデルを用いたものがある。

本研究ではネットワーク型モデルを用いたものを対象とする。その理由として、IRT では個々のテスト問題を独立したものとして扱い、その正誤で利用者の能力を評価するため、資格試験のように、すべて正答するためにはどうすべきかといった学習指針を示すことが困難だからである。本研究のように利用者の能力評価だけでなく、学習支援も目的とする場合、利用者ごとに行き詰まっている問題やその分野を把握し、それを解消するための指針をフィードバックすることが重要となる。

ネットワーク型モデルを用いた適応型テストは問題間の関係性をネットワークモデルとして仮定する。そして、その回答履歴や正誤情報より、優先して勉強すべき項目を提案するといった支援が可能であり、学習支援システムとして利用できると考えられる。

しかし、ネットワーク型モデルを用いた適応型テストの問題点として、事前に多数の問題を意味のあるカテゴリにまとめることへの難しさが考えられる。本研究で対

象としている試験は 100 の分野、及び「知識型」、「思考型」、「計算型」の 3 種類に分類できるという特徴をもつ。問題のテキスト化に加え、分野や種類といった属性を推定し、入力する作業は管理者にとって大きな負担となり得るため、これら作業の自動化が課題となっている。そこで、本研究では適応型テストにおける学習評価の前段階として、問題の種類別および分野別分類の精度向上を目的とする。

3. 問題文の分類

問題文などの文書を分類する手法として、主に確率モデル、ルールベース、機械学習を用いたものがある。ルールベースや確率モデルは、分類に有効な特徴を抽出するのが困難であることから機械学習を用いた分類手法について検討する。

また、本報告では、文書分類のように説明変数が多数になる分類問題に対してもうまく働く Random Forest を用いた分類手法について検討する。

3.1 Random Forest

Random Forest は複数の決定木を用い、識別を行う機械学習アルゴリズムである。Random Forest に用いられる個々の決定木は高い識別性能をもたないが、それらを複数用いてそれぞれの結果を補うことによって高い予測性能を得るアンサンブル学習ができるのが特徴である。

3.2 Random Forest の手順

本研究での Random Forest の手順について述べる。

① 文字列の抽出

本報告では問題文及びその正解選択肢を形態素解析エンジン MeCab^[4]により解析する。解析により抽出された形態素の有無を入力データとして用いる。

ここでは、分類に有効でない形態素を除くため品詞が名詞であるもののみを抽出対象とする。

② ブートストラップサンプルによるデータ集合の生成

ブートストラップサンプルとは N 個のサンプル集合の中から重複を許してランダムに標本を選んでできた新しいサンプル集合のことを指す。本報告では手順①ですべての問題から抽出した文字列に対して d 次元のデータ行列を生成する。そして、各問題における文字列の有無 $\{x_i\}$ をデータ行列 X によって表現し、データ行列 X と分野 C_i のセットを学習データとする。そして、これら学習データの集合からブートストラップサンプルにより学習に用いるデータ集合を生成する。

$$D_L = \{(X_1, C_1), (X_2, C_2), \dots, (X_N, C_K)\} \\ X = \{x_i\}_{i=1}^d, \quad \Omega = \{C_i\}_{i=1}^K \quad (1)$$

D_L : 学習データ集合, Ω : 分野の集合

C_i : 各分野, $x_i \in \{0,1\}$: 文字列の有無

X : 文字列の有無を特徴とした問題文のデータ行列

“Development of the Information Technology Learning Supporting System -A Study on the Classification of Problem Statement for Learning Evaluation-”

[†]Yusuke Miyakawa, [†]Takashi Izumi

[†]Nihon University

③ 決定木による学習

決定木の各ノード t において式(2)に示すジニ係数により学習に用いるデータ集合 D_t を2つの集合に分割する。ジニ係数は[0,1]の値をとり、値が大きいほど分割結果がばらついていることを示す。ジニ係数 I_G が最も小さくなるパラメータを推定し、入力データに対して{left, right}を返す分岐関数を $\{x_i\}$ ごとに作成する。そして、 $\{x_i\}$ から d 個ランダムに選択する。これを終了条件が満たされるまで続け、決定木を生成する。

$$I_G(t) = \sum_{i=1}^K P(C_i|t)(1 - P(C_i|t)) \quad (2)$$

$P(C_i|t)$: ノード t においてクラス C_i のデータが選ばれる確率

④ 複数の決定木による識別

手順②, ③により複数の決定木を生成し、図1に示すRandom Forestを構築する。これら決定木の出力はクラス C_i となっており、出力したクラスの多数決によって入力データの分類先が決定される。

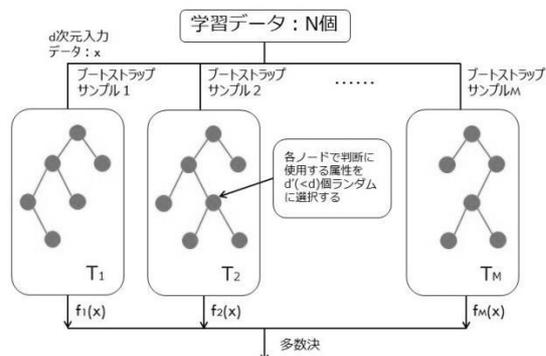


図1 Random Forestの構成

4. 実験

Random Forestを用いて問題文の種類別および分野別分類実験を行った。実験に用いる学習データやテストデータは表1, 表3の通りである。なお、分類正解率は全テストデータのうち、分野別分類に正解したテストデータ数を表しており、式(3)のように定義する。また、本実験ではPythonの機械学習ライブラリであるscikit-learn^[5]を用いた。

$$(\text{分類正解率}) = \frac{(\text{正解したテストデータ数})}{(\text{テストデータ数})} \times 100 [\%] \quad (3)$$

4.1 種類別分類実験

形態素を特徴量としたRandom Forestによる分類精度を検証するために、種類別分類実験を行う。表1に実験条件を示す。

表1 種類別分類実験における実験条件

| | |
|--------|---------------------|
| 種類別分類 | 平成22年度、20年度 春期・秋期 |
| 学習データ | 基本情報技術者試験 午前問題 320問 |
| 種類別分類 | 平成21年度 春期・秋期 |
| テストデータ | 基本情報技術者試験 午前問題 160問 |

表2 種類別分類正解率

| | 知識型 | 思考型 | 計算型 |
|----------|-------|-------|-------|
| 問題数 | 115 | 31 | 14 |
| 正分類数 | 115 | 10 | 11 |
| 分類正解率[%] | 100.0 | 32.26 | 78.57 |

表2の実験結果より知識型の分類正解率は100.0[%]となった。これは、文末に書かれている問い方等が同じである問題が多かったため、高精度となったと考える。しかし、思考型は32.26[%]と低い値となった。これは各問題に共通する形態素が極端に少ないことが原因であると考える。

4.2 分野別分類実験

種類別分類実験において、本手法は共通する形態素や単調な問い方である問題に対して良い分類精度であることが分かった。これを踏まえ、本手法の分野別分類における有効性について検証する。

表3 分野別分類実験における実験条件

| | |
|--------|---------------------|
| 分野別分類 | 平成20~23年度 春期・秋期 |
| 学習データ | 基本情報技術者試験 午前問題 640問 |
| 分野別分類 | 平成24年度 春期 |
| テストデータ | 基本情報技術者試験 午前問題 80問 |

表4 各階層における分類正解率

| | 小分類 | 中分類 | 大分類 |
|----------|-------|-------|-------|
| 分野数 | 100 | 23 | 9 |
| 正分類数 | 53 | 62 | 66 |
| 分類正解率[%] | 66.25 | 77.50 | 82.50 |

実験結果を表4に示す。小分類の分類精度は大分類に対し、16.25[%]低下した。これは、問題の分野が細分化するにつれ、大分類では有効であった特徴量が誤分類の要因になるためであると考える。

5. まとめ

本報告では、問題の種類別および分野別分類の精度向上を目的として、Random Forestを用いた分類手法について検討した。その結果、知識型では100.0[%]、大分類では82.50[%]の良好な結果となった。

今後は格文法等を用い、問題の内容を推定する分類手法について検討する。

参考文献

[1] 基本情報技術者試験 (レベル2) シラバス Ver3.0 : http://www.jitec.ipa.go.jp/1_13download/syllabus_fe_ver3_0.pdf (2014-12)

[2] 高橋暁子, 喜多敏博, 中野裕司, 合田美子, 鈴木克明: 「eラーニングにおける適応型テストの事例調査」, 教育システム情報学会全国大会講演論文集 36, pp.176-177 (2011-8)

[3] 平井有: 「はじめてのパターン認識」, 森北出版株式会社, pp.175-197, (2012-07-24).

[4] MeCab: Yet Another Part-of-Speech and Morphological Analyzer: <http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html> (2015-1)

[5] scikit-learn: machine learning in Python: <http://scikit-learn.org/stable> (2015-1)