

学習者データの統計分析により得られる学習者特性の考察*

大沢 泰貴[†] 大谷 康介[‡] 大金 夕騎[‡]
 森本 祥太[‡] 小島 大奨[‡] 松井 拓海[‡] 松田 健[‡]
[†]サイバー大学 IT 総合学部 [‡]静岡理工科大学総合情報学部

1 はじめに

授業評価アンケートの結果を分析し、分析結果から得られる情報を授業改善に活かすことや、学習者の成績と回答結果の関連性を分析するという研究は、従来から行われている [1]。一般的なアンケート分析では、選択式の設問から得られた定量的データの基本情報量を求め、数値データを用い分析するという手法がとられている。この手法では設問により問われている情報しか得ることができないという問題を抱えているため、自由記述の設問から得られた定性データをテキストマイニングの手法を用いて分析するという手法が用いられる場合もある [2]。しかしながら得られた分析結果から学習者特性を解釈することは人間の手に委ねられる側面が強いため、適切な評価を与えることは容易でない。

本研究では、自由記述文の形態素解析より得られた特徴単語群と学習者のテスト得点から特徴空間を形成し、テストの得点を元に特徴単語群をクラスタ化することで、得点傾向と特徴単語の関連性について仮説を立てる。そして得られ学習者の特徴空間を統計的手法により分析することで、仮説の妥当性を確認し学習者の特性を考察する。

2 形態素解析による特徴抽出

本研究で分析する自由記述のデータは、大学 1 年生を対象とした線形代数の授業で得られたものであり、期末テストの点数との紐付けが可能なのである。アンケートの設問は「大学の数学を勉強する上で困ったことを述べよ」という 1 つの項目のみで構成される。このただ 1 つの設問より得られた自由記述データと学生の期末テスト点数を分析し、得られた情報を本稿で示す。まず文章の特徴を明らかにするため、得られた文章集合

に頻出する単語群を形態素解析により抽出する。また対象とする単語は名詞に限定し、1 人の文章につき 1 回のカウントとする。形態素解析には RMeCab を使用する。結果を表 1 に示す。

表 1: 頻出単語群

単語	頻度	単語	頻度	単語	頻度
理解	25	線形	14	教科書	11
高校	19	公式	12	大学	10
勉強	16	代数	12	大変	10
授業	16	計算	11		
問題	14	数学	11		

分析のため、出現回数が 10 回以上の単語を対象に出現頻度順に $w_1, w_2 \dots w_{13}$ とラベル付けを行う。この特徴単語群を集合 W として、

$$W = \{w_1, w_2, \dots, w_{13}\}$$

とする。

3 特徴空間の定義

本章では前章で得られた特徴単語群を用い、学習者個人が持つ特徴空間について定義する。学習者 n が持つ特徴空間を $x^{(n)}$ として、

$$x^{(n)} = (\chi_n(w_1), \chi_n(w_2), \dots, \chi_n(w_{13}), S_n)$$

とする。このとき $\chi_n(w_i)$ は、学習者 n の自由記述文に単語 w_i が含まれていれば 1 とし、含まれていなければ 0 とするような指示関数とする。さらに S_n は学習者 n のテスト得点である。

例えば学習者 1 のテスト得点が 60 で、自由記述の文章が「線形代数を理解するには、問題演習が必要だと感じた。」である場合、学習者データ $x^{(1)}$ は

$$x^{(1)} = (1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 60)$$

となる。

*Consideration of Learner Characteristics Obtained by Statistical Analysis of Learner Data.

[†]Taiki Oosawa, Cyber University

[‡]Shizuoka, Yuki Oogane, Morimoto Syota, Kojima Daisuke, Matsui Takumi, Takeshi Matsuda, Shizuoka Institute of Science and Technology

4 特徴空間の統計分析

本稿では学習者の特徴空間から、学習者の得点と自由記述データに含まれる特徴単語群の関連性について調べていく。

まず特徴単語群を以下に従い2つの集合 W_u, W_d に分類する。

$$w_j \in W_u \quad \text{if} \quad \frac{\sum \chi_i(w_j) S_i}{\sum \chi_i(W_j)} \geq \frac{\sum S_i}{71}$$

$$w_j \in W_d \quad \text{if} \quad \frac{\sum \chi_i(w_j) S_i}{\sum \chi_i(W_j)} \leq \frac{\sum S_i}{71}$$

このとき集合 W_u, W_d は、

$W_u = \{ \text{理解, 問題, 勉強, 授業, 計算, 線形, 代数} \}$

$W_d = \{ \text{高校, 数学, 公式, 大学, 教科書, 大変} \}$

となった。上記ではある特徴単語が自由記述に含まれる学習者の平均点を全体の平均点と比べ、高ければ W_u 、低ければ W_d に分類した。これは得点傾向が高い学生ほど”理解”, ”問題” といった特徴単語が出現しやすく、得点傾向が低い学生ほど”高校”, ”数学” といった特徴単語が出現しやすい可能性を示唆している。このような得点傾向と特徴単語の関連を確認するため、前章で定義した特徴空間を主成分分析する。

主成分分析により得られた第1主成分を横軸、第2主成分を縦軸として、主成分負荷量を座標とした特徴単語群の散布図を図1に示す。得点はx、 W_u に含まれる特徴単語は○、 W_d に含まれる特徴記号は*で表示した。

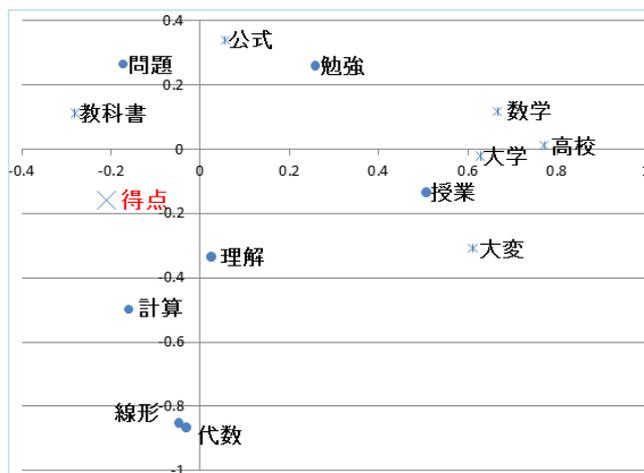


図 1: 散布図

図1の散布図において、集合 W_d に含まれる特徴単語”数学””高校””大学””大変”は密集している。さらに

得点を示す点と、第1主成分負荷量の正負が異なり、x軸(第1主成分)に関して比較的離れたところに存在している。

それらの特徴単語群同士の共起性と得点との相関性を確認するため、表2に相関行列を示す。特徴単語群は得点と負の相関を示しており、特徴単語同士では、正の相関を示した。

つまり本稿における分析では、得点傾向の低い学習者ほど、”数学””高校””大学””大変”といった特徴単語が、自由記述データに同時に出現しやすい傾向にあることを確認した。

表 2: 相関行列

	得点	高校	数学	大学	大変
得点	1.00	-0.13	-0.21	-0.06	-0.02
高校	-0.13	1.00	0.36	0.40	0.30
数学	-0.21	0.36	1.00	0.27	0.27
大学	-0.06	0.40	0.27	1.00	0.30
大変	-0.02	0.30	0.27	0.30	1.00

5 まとめと今後の課題

本研究では、学習者の得点傾向と自由記述データの関連性について、統計的手法を用い特徴単語群から考察を行った。結果として、得点傾向が低い学習者の自由記述に見られる特徴を確認した。実際に自由記述データを手作業で確認してみたところ、全体的に「高校数学と大学数学の差に苦労している」といった意味の表現が多く見られたが、得点傾向の低い学習者ではより顕著にみられた。

今後の課題として、得点傾向を段階別に分け、学習者全体に対して分析を行うことが挙げられる。また今回の特徴空間にテストの各問題の正誤を加味することで、より具体的な分析ができると考えられる。

参考文献

- [1] 松田健:類似文章抽出による自由記述分析と授業改善について、サイバー大学eラーニング研究第2号 pp.1-12 (2012)
- [2] 山西博之:教育・研究のための自由記述アンケート分析入門、外国語教育メディア学会、メソドロジー研究部会,pp.110-124 (2010)