N-029

データ分析手法の原理の理解に対する自信の分析

Analysis of confidence in understanding principles of data analysis methods

安田 健人† 島川 博光†

Kento Yasuda Hiromitsu Shimakawa

1. はじめに

近年、IT 業界だけでなく、製造や物流、医療などの幅広い業界においてデータサイエンティストの需要が高まってきている。しかし、データサイエンスの技術習得には、確率統計、最適化、プログラミングなど多様な知識、スキルが求められ、くじける者が多い。その要因として、理解に自信がないことが考えられる。

本研究ではデータ分析手法の理解度を推定する。くじける人の多くが、知識の整理、順序立てた説明ができないことに着目する。これらの人の多くが、ある箇所で思い違い、勘違いして自信を無くしていることが考えられるため、思い違い、勘違いしているかどうかの評価から自信の揺らぎを分析する。データ分析手法の原理の理解に対する自信を分析することで、データサイエンス教育のくじけを防止する。

2. データサイエンスに対する理解と自信

2.1 データサイエンス教育

データサイエンティストを育成するためには、形だけのリテラシーレベルの教育では成果は得られない。データサイエンティストには、数学的知識だけでなく、現場特有の用語や概念を理解する能力も求められる。データサイエンティストは、分析手法に基づく数学的根拠を理解し、その理解のもと、多様な分析手法から最適なものを選んで、眼前の実データに運用できる能力を持たなければならない。このようにデータサイエンティストには、広い分野の知識と、それを実問題に適用する問題解決能力が求められる。

2.2 自信と理解度の関係

自信がある人は対象分野の概念や手続きをよく理解しており、一方で自信がない人はこれらに対する理解が乏しいと考えられる。したがって、自信のある人は知識の整理、順序立てた説明ができる人が多く、自信のない人は何かの分野、概念について思い違い、勘違いしているため、知識の整理、順序立てた説明ができないと考えられる。

本研究では、どの分野、概念が理解できていないかを、記述問題に対する回答をテキスト分析することで判別する。理解できていない分野、概念を同定することで、くじけそうな学習者の支援を提案することができると考えられる。

† 立命館大学 Ritsumeikan University

3. 理解と自信の推定

3.1 全体の流れ

データを分析する課題を与える。被験者の解答からテキストデータを収集する。次に類似度の推定と理解度の分類をするために、正解例の単語の頻度で、理解しているかどうかを判別する。その判別には AnoGAN を用いる。判別後は、重要単語の特定をするために、正解例で出てくる単語を頻度の高い順に列挙する。最後に、勘違い、思い違いしているかどうかを評価することで自信の揺らぎを分析する。

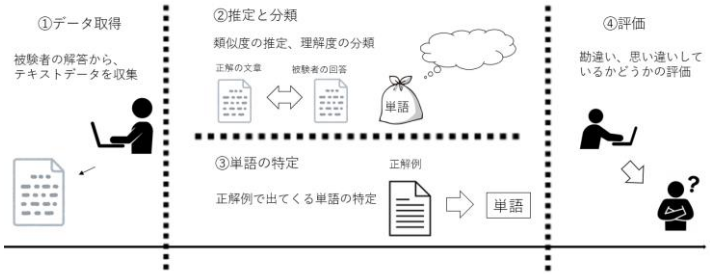

図 1 ：手法概要図

3.2 データ取得

データを分析する課題において、穴埋め課題、考察課題を出題する。本研究ではとくに後者に着目する。穴埋め課題では、データを分析するための Python コードに複数の空欄を設けて置き、この空欄を学習者が正しく補充できるかどうかを調べる。この課題から、学習者が分析のためのコードについての知識を正しく理解しているかを推定する。

考察課題では、上記の Python コードをデータに適用した結果、出力される数値が何を意味しているかを学習者に記述してもらう。考察課題から、学習者が、与えられた課題を分析するうえで指定された手法を使用する意義を理解しているか、かつ、その手法の適用結果を正しく読み解けていかを推定する。考察課題の回答文をテキストデータとして収集する。

3.3 類似度の推定と理解度の分類

学習者の回答と正解例との類似度は、dog2vec を使用して推定する。回答と正解例のそれぞれの単

語を、dog2vec を使用して分散表現に変換する。理解できない概念がある場合、回答における専門用語の頻度は、正解例のものとは類似性が低いと考えられる。この仮説に基づき、どの分野のどの概念が理解できていないかを導く。

図 2 は事前実験での回答例と正解例に出現する単語の類似度を dog2vec を使用して可視化している。事前実験では、大学での C 言語プログラミング演習での条件分岐を教えた単元の課題について回答と正答例の類似性を調べた。正解例で出てくる単語「else」、「return」に対して、回答に出現する「printf」、「認識」、「掲示」といった単語は正答するうえで必要性のない単語であり、必要性のある単語と必要性のない単語において距離があることが読み取れる。理解できていない学習者は必要な単語を使用せずに問題を解き、if-else 文の構文が理解できていないことがわかる。また、必要性のない単語から構文の使い方も理解できていないことが読み取れる。

本来出てくるはずのない単語の頻度で理解しているかどうかを推定するために AnoGAN を使用する。理解できていないときには、正しい解答の集合から大きく外れた、本来現れるはずのない単語が出てくることが考えられる。問題の解答に関係のない単語が多く含まれている分野、概念については思い違い、勘違いしていると考えられるため、理解が進んでいないと仮定する。

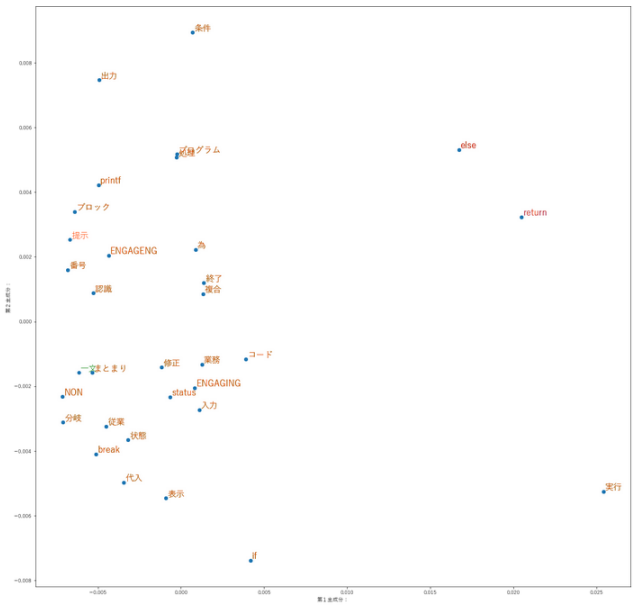

図 2：正解例に出てくる単語との類似度

3.4 重要単語の特定

正解例で出てくる単語を頻度の高い順に列挙する。正解例で出てくる頻度の高い単語が、学習者の回答に出現する頻度が高ければ、問題について理解が高い状態であると考えられる。正解例において出現頻度の高い専門用語が回答に含まれる確率を、ナイーブベイズを用いて算出する。また、理解している学習者は、回答文の文字数が多い傾向にあると考えられるため、文字数についても考慮に入れる。

3.5 評価

勘違い、思い違いしているかどうかを評価することで自信の揺らぎを調べる。くじけそうな人は、ある分野、概念について思い違い、勘違いをしていることで自信を失いつつあると仮定する。文字数や文量、専門用語の頻度、正解の文章を用意し、それとの類似度、本来出てくるはずのない単語の頻度で理解しているかどうかの分類についての評価をすることで、自信があるか、自信がないかを発見することができると考えている。

4. おわりに

本研究の結果を用いることで、人に教えられる人材を育成することができる。本研究は、知識の整理を補助する方法を発見することで、くじけそうな学習者を支援する。支援を受けた学習者は、知識の整理、順序立てた説明ができるようになる。これは、人に教えられる人材の育成へと繋がる。

参考文献

[1] 藤本佳久，“データサイエンス教育をめぐる状況の現状と課題”，人文科学論集　第 69 輯, 2023.

[2] 新濱 遼大,槇原 絵里奈,小野 景子,大正 歩夢,“オンラインジャッジシステムの問題選択支援 に向けた問題文間の意味的類似度の算出”,第 29 回ソフトウェア工学の基礎ワークショップ (FORSE2022),p43～48

[3] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, “Efficient Estimation of Word Representations in Vector Space”,International Conference on Learning Representatins(2013),arXiv preprint arXiv:1301.3781

[4] Ryuji ODA , Tsuyoshi MIKAMI, Yusuke SHINOZAKI, Kentaro KOSAKA , Masanao SAT, “SOM-based Feature Map for Searching Similar Japanese Materials used in Junior/Senior High Schools”, ファジィシステムシンポジウム 講演論文集,第 37 回(2021)

[5] Hiroki Unno,“Consideration of Improving Anomaly Detection Techniques Using Generative Adversarial Network”, The 36th Annual Conference of the Japanese Society for Artificial Intelligence, 2022

[6] Satoshi Shimizu,Yoshiaki Mizunochi、A case of study on the effects of teachers’ guidance by grasping learners understanding and confidence in advance, 日本科学教育学会研究報告 Vol34 No.5(2020)

[7] Kazuhisa Miwa,Tsutomu Shimomura, Definition of Understanding Potential Score Based on S-P Chart and Its Experimental Discussion, Electronics and Communications in Japan,Volume77,Issue6,p1～10

[8] 辻智,「教室の風景 先駆けとしての成城大学のデータサイエンス教育」，成城学園研究所 編『成城教育』第 195 号, 2022