高校生向けデータサイエンス教材の提案と使用ログの解析

村上綾菜 (指導教員：伊藤貴之)

1　はじめに

政府が AI 戦略を掲げ，高校生の段階から AI やデータサイエンスの基礎を学ぶよう，提言されている．しかし，既存の学習教材は，初学者にとって難解なものが多い．本研究では，高校生を対象としたインタラクティブなデータサイエンス教材の一事例を提案する．学習内容の一例として判別分析を採用しており，生徒は訓練データのクレンジング作業を通じて判別分析の仕組みを理解することを期待している．本研究では，本教材を授業において使用した際の使用ログを収集し，解析した．

2　関連研究

データサイエンス手法の学習を目的とした既存の学習教材は大きく分けて 2 種類に分類できる．1 種類目は，Web サイトを利用したインタラクティブな学習ツールである．具体的には，科学の道具箱や Bowland Japan が公開している教材が挙げられる．これらの教材の利点は，パソコン操作を苦手とする生徒でも楽しく学ぶことができる点である．一方で，現在このようなインタラクティブな教材で高校生を対象とした事例はまだ少なく，充実しているとはいいがたい．

2 種類目として，分析ソフトやプログラミング言語を用いた本格的なデータ分析環境があげられる．文部科学省が提示する高等学校情報科の教員研修用教材においては Excel を活用した単回帰分析や，Python や R を用いた分析の例がソースコードとともに紹介されている．神部ら [1] の研究では，R を用いた統計教育への問題解決型教教材を採用しているが，機器の操作方法の習得へ意識が奪われている学生がかなり見受けられたと報告されている．これらの教材の課題は，プログラミング初学者には苦手意識が先行し，データサイエンス自体も難解に感じる可能性があることである．その一方で，実践レベルの専門的な分析手法が学べるという利点もあり，一定以上の学力と PC スキルを有する高校生の学習レベルには適している面がある．

3　データサイエンス教材の開発

本章では我々が開発したデータサイエンス教材の概要と詳細について説明する．我々は R の Web アプリケーション作成パッケージである shiny を用いて本教材を実装した．本教材を Web アプリケーションにて実装した理由は，本教材使用者が事前のインストール作業等の煩雑な環境構築をすることなく使用できるようにするためである．

我々が開発した Web アプリケーション教材のメイン画面のスナップショットを図 1 に示す．この教材において，画面左側には，判別分析に使用するデータを散布図で表示する．画面右側には，ユーザが使用する各種機能に関するボタン，および判別分析の精度表示のための GUI 部品が搭載されている．

図 1: 教材メイン画面のスナップショット．

3.1　高校生向けの教材としての工夫

我々は，本教材の学習目標を以下の 3 点とした．

目標 1: 生徒が高校生の学力の範囲内で判別分析の概念を理解できる．

目標 2: プログラミングや複雑な操作ができなくても，生徒が教材を使いこなせる．

目標 3: 生徒が楽しくデータサイエンスを学習できる．

我々は，この学習目標の達成を目標として本教材を開発する．同時に，高校生が使用することを鑑み，本ツールには主に 4 点の工夫を施すこととした．

1 点目として，**目標 1** のために，データの視覚表現を 2 次元の散布図に限定する．これには，高校生が散布図を理解しやすくする目的がある．高校生は数学 I データの分析の章で「散布図や相関係数を用いて 2 つのデータの相関を把握し説明すること」を学ぶ．つまり，高校生は 2 次元の散布図は見慣れているが，変数が 3 つ以上のグラフは見慣れていない．見慣れないグラフは負担になりやすいため，今回は使用するデータの変数を 2 つに制限した．

2 点目として，**目標 2** のために，生徒の操作に対して即時のフィードバックを与える．具体的には，散布図上の点を削除するたびに，散布図上で判別分析の境界線が動く．この動きにより，作業のたびに自分の動作が正しいかどうかを確認できるので，試行錯誤して課題を遂行することが期待できる．

3 点目として，**目標 2** のために，全ての操作を GUI ベースにし，クリック操作のみでほぼ全ての作業を行えるようにする．GUI の工夫により，生徒にとって負担になる可能性があるキーボード入力を排除する．これには，学習内容以外の負担を減らすことで学習内容へ集中を促す目的がある．

4 点目として，**目標 3** のために，高校生にとって身近なデータを題材とする．今回は J-POP 等の楽曲を題材とした．データの中に自分の知っている曲を見つけることで，データが身近であることに気がつき，同時にデータサイエンスも身近であることを理解することを期待する．

3.2 使用したデータ

本教材の判別分析に使用したデータは，600 曲以上の J-POP の楽曲の音響特徴量である．具体的には，Librosa を用いて楽曲の音響信号から音響特徴量を算出し，これを説明変数とした．我々は，高校生がデータを理解しやすいように，音響特徴量に対して次元削減を適用して 2 次元に圧縮し，それを座標値にして散布図として表示した．次元削減には UMAP を使用した．判別のラベルは楽曲の発売年を適用し，手作業で付与した．今回は楽曲の発売年を「1999 年以前」と「2000 年以降」で 2 種類に分類してラベルとした．

3.3 学習教材

本教材においてユーザに課する作業は「散布図上の点をクリックすることによる例外的なデータの削除」であり，その結果としてユーザに求める目標を「分類精度の向上」であるとした．この作業をユーザが円滑に遂行できるように，本教材では以下の機能を実装する．

- 分類精度の確認
- 判別の境界線の確認
- 散布図上のデータの詳細情報の確認
- 削除したデータの一覧表示の確認
- テストデータプロット位置の確認

ユーザは上記の機能を適切に使用しながら，判別分析の精度を向上するためにどの点を削除すべきかを考えるものとする．

3.3.1 メイン画面

図 1 に示したメイン画面は，ユーザがデータクレンジング作業を遂行するための画面である．この画面の散布図上で，ユーザは例外と思われる点をクリック操作によって削除する．散布図には判別の境界線が直線で表示されている．点を削除するたびに本教材では判別分析を再度実行し，その結果として散布図上の境界線が変動する．マウスのカーソルを散布図上の点に重ねると，その点に対応する楽曲の詳細情報が散布図の下に表示される．ユーザが例外的な点を削除する際には，散布図上のプロット位置からだけでなく，楽曲の詳細情報も同時に読んで総合的に判断することもできる．

判別精度を評価するために，我々は，楽曲データ全体をあらかじめ訓練データとテストデータに分割した．本教材では，判別分析による境界線決定のための学習に訓練データを使用し，判別精度の検証のためにテストデータを使用する．

3.3.2 テストデータ確認画面

テストデータ確認画面で，ユーザはテストデータのプロット位置と境界線の位置関係を確認することができる．これにより，ユーザは精度算出に用いられている該当テストデータが，散布図上のどの点に対応するのかを確認できる．また，メイン画面と同様に，散布図上の点にマウスオーバすることで，その点に対応する楽曲の詳細情報を読むことができる．

3.3.3 削除データの一覧表示画面

削除した点の一覧表示画面では，ユーザが削除した点の詳細情報が一覧表示される．これを適切に使用することで，学生が削除したデータの傾向を理解し報告する，という発展課題を設定することも考えられる．

4 操作ログの解析

本研究では，お茶の水女子大学理学部情報科学科の学部 2,3 年生合計 26 名を被験者として本ツールを用いて学習してもらい，その操作ログを記録した．操作ログでは，被験者が所定の操作を行うごとに，その時刻と操作内容を記録する．ここでは記録した操作ログのうち，削除したデータに注目する．

散布図の点の濃さでデータの削除順序を表現し，学生ごとに可視化する．その可視化結果から，学生の削除傾向を 3 つのグループに分類できた．散布図の両側から均等に削除する者，散布図のどちらか片側から集中的に削除し始める者，散布図中央を中心に削除する者，の 3 つである．また，削除傾向と作業時間と照合することで，各学生の作業効率を推定することができた．可視化結果および解析結果の詳細は，他の文献 [2] に掲載している．

5 まとめと今後の展望

本研究では，高校生を主対象として，判別分析を例題とした Web アプリケーション型のデータサイエンス教材を提案した．我々は，データサイエンス初学者に本教材を使用してもらい，その使用ログを解析した．この結果，本教材の使用者は，学生ごとの傾向はあるものの外れ値の概念を理解し，判別分析の境界線から離れた点を優先的に削除することを発見した．

今後の課題として，ユーザ全体を分類精度の向上を効率的に達成したグループとそうでないグループに分類し，その違いに着目したい．Shirvani ら [3] の研究では，ページ遷移を類似した手順でクラスタリングし各クラスターの学習特徴を分析する．本教材においても，3 つのページが存在するので，ページ推移と学習効果の関係にも着目し，さらなる解析を進めたい．

謝辞

音楽データをご提供くださった株式会社レコチョク社の関係者の皆様に感謝の意を表します．

本教材を授業で使用してくださったお茶の水女子大学附属高等学校の先生方に感謝の意を表します．

本研究の一部は，日本学術振興会科学研究費補助金の助成に関するものです．

参考文献

[1] 神部順子，玉田和恵，“ プログラミング活用による統計教育への問題解決型教材の開発に向けて ”，Informatio: 江戸川大学の情報教育と環境，vol. 17，pp. 29-32，2020.

[2] 村上綾菜，伊藤貴之，“ 高校生向けデータサイエンス教材の開発と使用ログの解析 ”，DEIM，2021.

[3] Boroujeni Mina Shirvani, Pierre Dillenbourg, “ Discovery and Temporal Analysis of Latent Study Patterns in MOOC Interaction Sequences ”, LAK '18: International Conference on Learning Analytics and Knowledge, pp. 206-215, 2018.