

【研究報告】

長瀬 永遠

富山県立大学 情報基盤工学講座

December 1, 2023

背景

近年、コンピューターの普及によって社会にはあらゆる分野において大量の情報が溢れるようになった。そのため、大量のデータから有用な情報を取り出す研究が数多くなされている。その中の一つに複数の入力データと単一の出力データからそれらの間に成り立つ数法則を発見する手法が挙げられる。

目的

数法則発見法を社会における事象に適用するために潜在プロファイル分析の考え方を加えた手法を提案する。

新しい政策の効果検証がしたい…

理想

- 予算 ∞
- すぐに適用できる
- すぐに効果が出る

現実

- 予算簡単につかない
- 適用に数年
- 効果が出るのに数年

疑似的に効果検証できるようにすればよいのでは？

➡ そのためには社会を表現したモデルが必要

数法則発見法

以下の二つを検討している.

RF5.0

複数の入力変数（量的）と一つの出力変数（量的）との間に成り立つ数法則を求めることができる 3 層パーセプトロンを用いた多変量多項式回帰法

RF6.3

RF5.0 の考え方をもとに入力変数のみ質的変数も適用可能に拡張した手法

疑問

日本全国を見てみると人口や主要産業、地価など大きなばらつきがあるが、全てを同じ土俵で扱っていいのか？

仮説

- 全国の市区町村は潜在的にいくつかのタイプに分かれるのではないか
- タイプ分けすることで数法則発見の精度が向上するのではないか

アイデア

多数の入力変数（量的）を用いてクラスタリングを行い、発生したクラスを質的変数として入力変数に組み込むことで RF6.3 を適用する.

クラスタリング手法の要件

- クラス数を事前に定義する必要がない
- ソフトクラスタリングである

クラスタリング手法

上記の要件を満たす手法として潜在クラス分析（前回の研究会で紹介）をもとに連続変数に対して適用可能に拡張された潜在プロファイル分析を用いることを検討中。

動作確認 1

コードが正常に動作しているか確認するために以下のデータを使用して分析を行った.

- セ・リーグ（プロ野球）2023 年シーズンにおける打撃成績
- 条件：50 打席以上が記録されている選手
- データ数：107
- 項目：打率，出場試合数，盗塁，OPS. 等を含む 18 項目
- 参照元：プロ野球データ Freak より収集

動作確認 2

BIC を用いて適切な潜在クラス数を求めた結果は以下のようになった．この例ではクラスを 6 つに分けることが適切と判断できる（このパッケージでは符号が反転するためより大きい値をとるクラスが良いとされる）

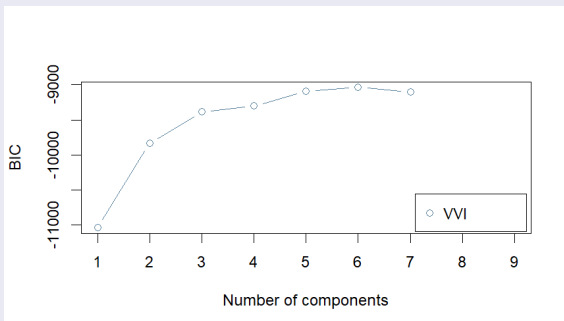


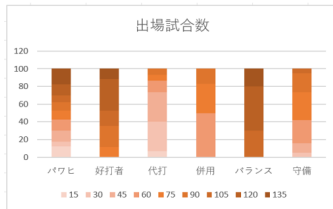
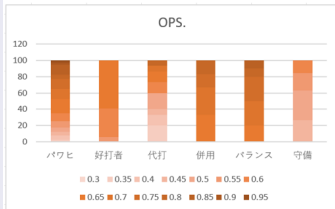
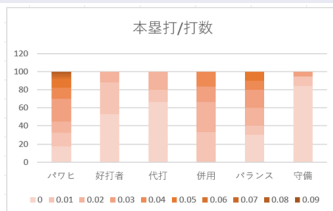
図 1: 潜在クラス数ごとの BIC

depmixS4 による潜在プロファイル分析の実装3

9/10

動作確認 3

各クラスに属する選手の特徴をつかむために打率，1打数当たりの平均本塁打数，OPS.，出場試合数の4項目についてクラスごとの傾向を図示した。



現状考えている方法

- 1 データにつき 1 クラスとして最も所属確率が高いクラスを用いる
- いくつかのパターンで RF6 を行い、最も妥当性の高いものを採用する

- 潜在プロファイル分析の出力を RF6.3 の入力に加える手法の検討
- 効果検証の方法とデータを検討
- 修士論文の執筆