

カテゴリカルデータに対する 解釈しやすいクラスター視覚化法の提案

同志社大学大学院文化情報学研究科 高岸 茉莉子

Erasmus University Rotterdam Michel van de Velden

多変量カテゴリカル変量間の関係を視覚的に解釈するための方法として多重対応分析 (Multiple Correspondence Analysis, MCA) がある。MCA では低次元空間上の対象、カテゴリの座標が推定され、座標間の距離に基づき関連の強さを解釈する。このような図はパイプロットとも呼ばれる。パイプロットに外部情報 (座標の推定には用いないが、パイプロットにそれに関する解釈を加えたい場合に使う情報) を加えると、解釈がより行いやすくなる。本発表では、外部情報として性別などのカテゴリカル変量を考える。そして対象を外部情報ごとにクラス分けし、対象とカテゴリの関係の、クラスごとの特徴を視覚的に解釈することを目的とする。

上記目的を達成する単純な方法として、対象の座標をクラスごとに平均をとり、それを1つの座標点とする方法が考えられる。しかしこの方法ではクラス内の大多数の人が同じ傾向を持つ場合は、その傾向はパイプロット上でも解釈しやすいが、そうでない場合は解釈が難しくなる。例えばクラス内で複数の傾向に等分に分かれている、また同じ傾向を持つ対象の中でも少数派のクラスが含まれている (例、ある傾向を持つのは大半が若者だが、多少の高齢年齢層もいる)、などの状況は、外部情報ごとの特徴を知る上では有益な情報になりうるが、平均をとる方法では視覚的に解釈することは難しい。

そこで本報告では、上記のように外部情報クラス内で複数の異なる傾向がある場合、これらの傾向を1つのパイプロット上で解釈するための方法を提案する。具体的には、外部情報クラスごとに複数のクラスターを抽出し、それら全てを共通の低次元空間上に同時布置する。ここでクラスタリングすることは、各クラス内で似た傾向を持つ対象のみに人数を絞った上での、相対的なカテゴリとの関連の強さを見ることを意味する。これにより、例え少人数のみが持つ傾向であっても、関連の強さの情報が保たれ、視覚化結果にも反映されやすくなる。また異なる外部情報クラスのクラスターを全て同じ空間上に布置することで、外部情報として複数の変量を用いることも可能となり、更に異なる外部情報クラス間の関係も視覚的に解釈できる。

また MCA に対象の外部情報を加える手法の関連既存研究として、外部情報を加えることを線形制約を課すアプローチとして一般化した [1] があるが、本報告では、提案手法がクラスター対応分析 (クラスタリングと MCA の同時分析 [2]) に線形制約を課すアプローチの特殊ケースとして捉えられることも示す。更に当日は数値例や実データ例を通じて、本手法の有用性を示す。

参考文献

- [1] Hwang, H., & Takane, Y. (2002). Generalized constrained multiple correspondence analysis. *Psychometrika*, **67**(2), pp.211–224.
- [2] van de Velden, M., D'Enza, A. I., & Palumbo, F. (2017). Cluster Correspondence Analysis. *Psychometrika*, **82**(1), pp.158–185.