統計数理研究所 75 周年記念事業（2019）

集約的シンボリックデータ解析

Aggregated Symbolic Data Analysis

データ科学研究系　清水信夫（Nobuo Shimizu）

近年，様々な分野において，Web システムを用いたデータの収集が多用されており，各分野における活動の詳細なデータが計算機上に連続的に蓄積されるようになっている．それらのデータは連続変数とカテゴリー変数が混在した多次元データであることが多く，データの個体数についても非常に大きな場合が多々存在する．

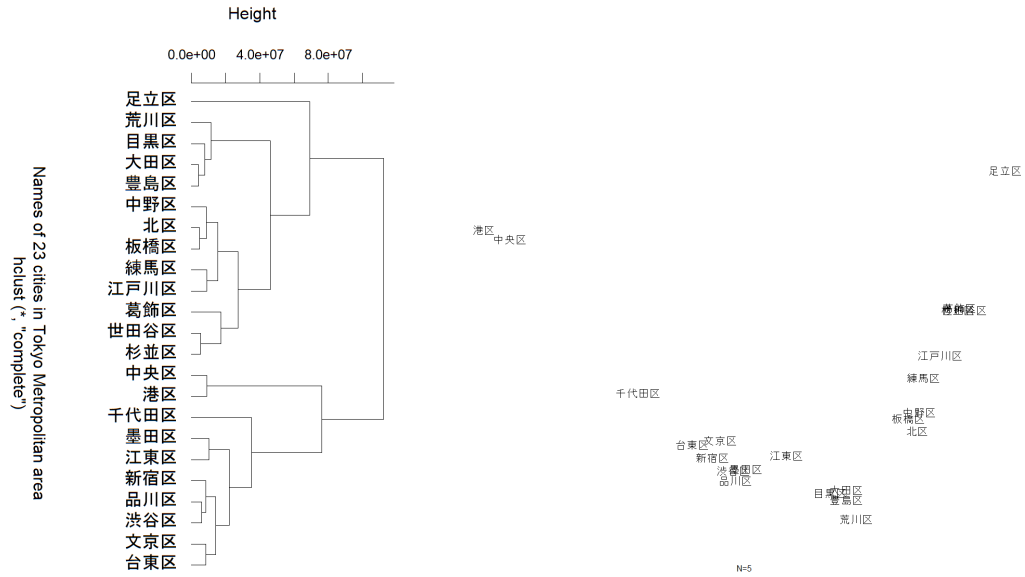
このようなデータは，全体像を見たり詳しい解析を行ったりするための，一般的なデータ管理・処理ソフトウェアによる取り扱いが困難なほどの巨大かつ複雑なデータ，いわゆる“ビッグデータ”の代表例と考えられ，従来多用されていた個体データに着目する方法以外の解析手法の開発が求められる．ただし，そのようなときは，個体データが意味のある自然な比較的少数のグループに分かれる場合が少なからず存在する．したがって，個体データそのものではなく，個体がまとめられたグループに関心を向けた手法が必要である．その方法の一つとして，Diday(1988) によりシンボリックデータ（Symbolic Data, SD）という概念が提案され，データの中で個体がまとめられたグループを SD として解析するシンボリックデータ解析（Symbolic Data Analysis, SDA）が提唱されている．

SDA においては，データとして各連続変数ごとに 1 つの値ではなく，ある値を中心としてばらつきをもつデータ（区間データや分布値データ）などで表されるものが考えられ，それらを SD と考えた場合の解析として従来の各種多変量解析手法を拡張する研究が，Bock and Diday(2000)，Billard and Diday(2005)，Diday and Noirhomme-Fraiture(2008) などにまとめられている．それら以外にも，SD に関するクラスタリングに関しては Verde(2004) や Irpino and Verde(2006) など，多次元尺度構成法に関しても Denœux and Masson(2000) や Groenen et al.(2006) などの研究がある．これまでの SDA においては，データは最初から区間や集合のような形で与えられている場合が多く，そこではグループ内の複数の変数間の関係は無視される．例えば，2 つの連続変数間の相関関係は考慮されない．

しかしながら，現代のビッグデータにおいては、元の個体データは保持されている．個体数や変数が極めて多いデータの場合は移動や計算に困難を伴うが，どうしても必要ならばグループに関するいかなる記述統計量も計算することは可能である．そこで，グループにおける多次元データの情報を可能な限り簡潔な形で持つために，複数の記述統計量を考えることにし，それを集約的シンボリックデータ（Aggregated symbolic data，ASD）と呼ぶこととする．

ASD は，グループ内の個体データのそれぞれの変数および複数の変数の組み合わせに関し，情報の脱落を可能な限り抑えつつ保持すべき値を可能な限り少なくして取り扱いを容易にするために，2 次までのモーメントについて求められる統計量の集合として表される．ASD に含まれる統計量の例としては連続変数における平均および分散共分散行列，カテゴリー変数における分割表などがある．これらを用いて，ASD 間の非類似度を尤度比検定統計量やカイ 2 乗統計量などを用いて定義し，それらを用いてクラスタリングや多次元尺度構成法を行う方法を提案した．この方法を東京都区部の不動産情報データに適用し，23 の特別区をそれぞれグループとして考え，データの各変数から各グループ間の非類似度を ASD を用いて求めた上で階層的ク

ラスタリングおよび多次元尺度構成法を行った例を図 1 に示す.



(a) 階層的クラスタリング　　(b) 多次元尺度構成法

図 1. 集約的シンボリックデータを用いた不動産情報データの解析例
(清水・中野・山本 (2018))

参 考 文 献

Billard, L. and Diday, E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*, John Wiley & Sons Ltd, Chichester, UK.

Bock, H.-H. and Diday, E. (2000). *Analysis of symbolic data: exploratory methods for extracting statistical information from complex data*, Springer-Verlag, Berlin.

Denœux, T., Masson, M. (2000). Multidimensional scaling of interval-valued dissimilarity data, *Pattern Recognition Letters*, **21**, (1), 83–92.

Diday, E. (1988). The symbolic approach in clustering and related methods of data analysis, *Classification and Related Methods of Data Analysis*, 673–684.

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*, John Wiley & Sons Ltd, Chichester, UK.

Groenen, P. J. F., Winsberg, S., Rodriguez, O. and Diday, E. (2006). I-Scal: Multidimensional scaling of interval dissimilarities, *Computational Statistics and Data Analysis*, Elsevier, **51**, (1), 360–378.

Irpino, A. and Verde, R. (2006). A new Wasserstein based distance for the hierarchical clustering of histogram symbolic data, *Data Science and Classification*, Springer, Berlin, 185–192.

Verde, R. (2004). Clustering methods in symbolic data analysis, *Data Science and Classification*, Springer, Berlin, 299–317.

清水信夫, 中野純司, 山本由和 (2018). 集約的シンボリックデータのカイ 2 乗統計量を用いた非類似度とその不動産情報データへの適用, 統計数理, **66** (2), 279–294.