区間値関数データの解析

データ科学研究系　計算機統計グループ
助教　　清水　信夫

1　はじめに

科学一般もしくは人間の社会的活動において、データを解析した上で、そこから何らかの情報を得ることは普遍的な作業として広く行われている。ここでデータとして想定するものは、解析の目的や利用可能な環境に依存するが、従来の統計学においては各々の個体が単一の量的データもしくは質的データをとる場合を前提とする議論が多く、それらを数値化した上でベクトルや行列として表現する手法が発展してきた。これらの概念や表記法は有用ではあるが、近年ではこれらの枠組みでは表現できない各種データ（例として時系列データ・空間データ・時空間データなど）を解析する手法の開発が統計学において期待されている。この新たな期待に応える方法論として、関数データ解析やシンボリックデータ解析があり、両者とも1980年代に提案されて以降、多様な研究が現在に至るまで活発に行われている。本報告ではこの両者の特徴を踏まえ、区間値関数データの解析について述べる。

2　関数データの解析

解析対象とする個体がそれぞれ系統的に観測されたデータをもつとき、それらを関数化したものを新たなデータ（＝関数データ）とみなして解析を行う考え方が、Ramsay(1982)により提案された関数データ解析である。関数データ解析においては、最初から関数データが与えられている場合も含め、関数データの微分なども利用可能という点で従来の統計的データ解析よりも多様な解析手法を考えられるという特徴がある。また、個々の関数データを有限個の基底関数を用いて近似展開して有限次元の多次元データに変換し、それらに対して従来型の各種解析手法を適用することもできる。更に別のアプローチとして、変分法を利用することで関数データを無限次元のまま解析する方法も研究されている。関数データ解析についてはRamsay & Silverman(1997,2005)などの成書が参考になる。

関数データ解析の中でも活発に研究が行われている手法として関数クラスター分析法がある（水田(2003,2004)、Mizuta(2002,2003)など）。そのうち、関数データの定義域に依存しない手法においては、有限個の基底関数により近似展開された個々の関数データ間の距離や非類似度を計算することができ、それらの値を用いて各種階層的手法・*k*-means法・主要点法などを適用することができる。

3　シンボリックデータの解析

新しいデータの概念として、Diday(1987)などにより提案されたシンボリックデータがある。これは従来の統計学において広く利用されてきたデータ型の枠組みを一般化して、多様なタイプのデータを許容する形で定義されたものであり、またそれらを解析する手法がシンボリックデータ解析である。シンボリックデータのオブジェクトは単一の量的データもしくは質的データのみならず集合データ、区間データ、分布なども含むことができ、またそれらに重みをつけることも可能である。更に、シンボリックデータを解析した結果（クラスター分析における各クラスターなど）自体もオブジェクトとして扱える。これはすなわち、データが階層構造を有する場合についても扱えるということを意味する。シンボリックデータ解析を行うソフトウェアとしては、SODAS（Symbolic Official Data Analysis System）が公開されている。Bock & Diday eds.(2000)、Billard & Diday(2006)、Diday & Noirhomme-Fraiture eds. (2008)などは、シンボリックデータ解析や

SODASに関する成書として参考になる。

これらの多様なシンボリックデータのうち、最近では特に区間データに対する様々な解析法が研究されている。区間データの解析については、各次元の区間の最小値および最大値、もしくは区間データの重心の値を用いることで、従来型のデータの解析と同様の問題に帰着させる方法が多く提案されている。また、区間データ全体について成立する解を求める手法についても研究が行われている。

4　区間値関数データの解析

区間値関数データとは、関数データの定義域内のそれぞれの値に対応する値域が区間データとして表されるもののことであり、それぞれの値域の最小値を結ぶことで得られる関数（下界関数）と、最大値を結ぶことで得られる関数（上界関数）により幅をもつ関数データとして表される。

区間値関数データの解析については、関数データ解析法とシンボリックデータ解析法における区間データの各種解析手法を併用することにより、各個体のデータの下界関数および上界関数（および平均関数）の値を用いて行うことができる。一例として、定義域に依存しない区間値関数データのクラスター分析を考える。この手法においては、各個体間の距離を区間データの解析における様々な距離基準（Hausdorff距離、Gowda-Diday非類似度など）の拡張の形で求めることができ、それを用いて従来型のデータの場合と同様に解析を行うことができる。また、データの値によっては関数データ解析における有限個の基底関数による近似展開を行うことで、有限次元の多次元データのクラスター分析を行う問題に帰着させることも可能である。

5　おわりに

従来型のデータ構造では記述できない新たなデータ型の例として区間値関数データを取り上げ、それを解析する手法についてクラスター分析を例にとり報告した。今回の報告内容は、「多様なデータ構造に対応可能な統計的データ解析手法の開発」の第一歩であり、従来の統計学の枠組みのみならず、それらに関連する諸科学や情報科学などとの研究交流により更なる発展の可能性が考えられる。

参考文献

Billard,L. and Diday,E.(2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley.

Bock,H.H. and Diday,E. eds.(2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Springer-Verlag, Berlin.

Diday,E.(1987). The symbolic approach in clustering and related methods of data analysis, *Classification and Related Methods of Data Analysis, Proceedings of IFCS '87, Aachen, Germany*.

Diday,E. and Noirhomme-Fraiture,M. eds.(2008). *Symbolic Data Analysis and the SODAS Software*, Wiley.

Mizuta,M.(2002). Cluster analysis for functional data, *Proceedings of the 4th Conference of the Asian Regional Section of the International Association for Statistical Computing,* 219-221.

Mizuta,M.(2003). K-means method for functional data, *Bulletin of International Statistical Institute, 54th Session,* Book 2, 69-71.

水田正弘(2003). 関数データに対するクラスター分析, 2003 年度統計関連学会連合大会講演報告集, 429-430.

水田正弘(2004). 超高次元データとしての関数クラスター分析, 2004 年度統計関連学会連合大会講演報告集, 365-366.

Ramsay,J.O.(1982). When the data are functions, *Psychometrika*, **47**, 379-396.

Ramsay,J.O. and Silverman,B.W.(1997). *Functional Data Analysis,* Springer-Verlag, New York.

Ramsay,J.O. and Silverman,B.W.(2005). *Functional Data Analysis 2nd Ed.* Springer-Verlag, New York.

清水信夫(2009). 区間値関数データのクラスター分析について, 2009 年度統計関連学会連合大会講演報告集, 98.