解 説

データの形が教えてくれること
—トポロジカル・データ・アナリシスとその応用—

梅田裕平（（株）富士通研究所）

データには形がある

近年，ビッグデータやモノのインターネット（Internet of Things / IoT）といった言葉が世間を騒がせている．大量のデータを集め，扱うことができるようになったことで，さまざまなことができるようになってきた．また，ディープラーニング（深層学習）をはじめとする大量のデータ収集が課題だった人工知能技術の発展にも大きな影響を与えている．その一方で，ビッグデータやIoTが成功するか否かはデータ分析にかかっているといわれている．というのも，データは単に集めるだけでは意味がなく，実際に応用に繋げるためには，そこから何かしらの「知見」を取り出す必要があるためである．現状のデータ分析に関する技術はビッグデータ以前からの統計的な手法をもとにしたものが主に使われているが，データが大量かつ複雑になっていく中で，従来のやり方だけでは詳細な分析が難しくなってきており，重要な情報を見逃すことも多くなっている．そういった中で近年新しいデータ分析手法として注目され始めているのが，データの形を捉えることで新たな知見を得ようとするトポロジカル・データ・アナリシス（Topological Data Analysis / TDA）である．本稿では，データの形をどのようにして捉えるのか，それによってどういったことが分かるのかを応用事例とともに解説する．

データの形を見るということ
—トポロジカル・データ・アナリシス—

図-1と**図-2**を見てみよう．これらは，2次元のデータの分布を図示したものであるが，明らかに違う特徴を持っていると分かるのではないだろうか．

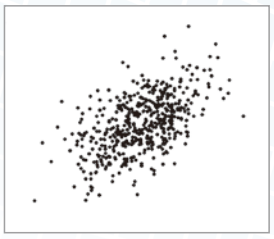

図-1　2次元データの分布の例

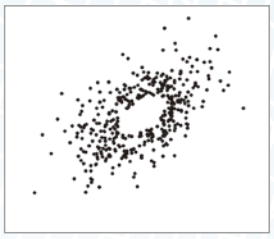

図-2　穴の開いた2次元データの例

図-1と違い，図-2にはデータの中央に穴が開いている．これらのデータを見た場合，それぞれどのようにデータの解析をしようと思うか考えてほしい．図-2のデータでは穴の周りのデータがどうなっているか調べようと思う人も多いのではないだろうか．実際，穴つまりデータが発生しない状況は，設定の不具合であったり機器の運転の制約であったりと重要な情報になることが多く，このような部分の解析は慎重に行われていることが多い．

一方で図-1と図-2を一般的なデータ解析手法で解析することを考えてみよう．たとえば，データ解析の常套手段である主成分分析をしてみる．すると，この2つの主成分の結果はほぼ同じ結果となる．その結果，この2つのデータはほぼ同じものと捉えて，穴の周りのデータを調べようとは思わないのではないだろうか．穴の周りのデータを調べようと思いつくのは，データを2次元で図示し，目で見て形を捉えることによってできることであり，データの形を捉えるということの重要性を表している例である．

ところで，図-1と図-2は2次元のデータであったので図示することができたが，ビッグデータやIoTの時代と言われる中では3次元以上のデータが中心となり，目で見て形を捉えることは不可能になってくる．このような場合，どのようにデータの形を捉えたらいいだろうか．

その解決策の1つとして考えられているのが数学の一分野である幾何学の手法を使うことである．幾何

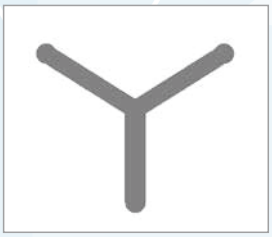

図 -3　Y 字型の図形

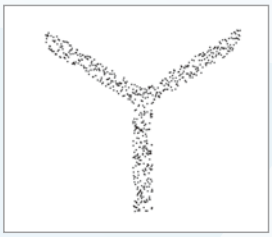

図 -4　Y 字型に分布したデータ

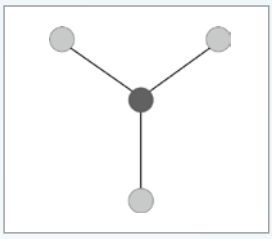

図 -5
Mapper に Y 字型に分布したデータを入力した場合の出力例

学はもともと高次元を含めた（一般化された）図形を数値や数式で把握する学問であるため，その技術をデータに応用すれば，データの形を数値や数式で捉えることができると考えられる．TDA は，幾何学の一分野である位相幾何学（トポロジ）の手法を使ってデータの形を捉えながらデータを解析する手法であり，従来の手法では把握できなかったデータの知見を取り出そうという技術である（文献 1）を参照）．

以降の章では，TDA の鍵となる 2 つの手法について簡単に説明し，それらがどのようにデータ解析に応用されているか紹介したい．

データのどこに注目するか

❖ モース理論と Mapper

図 -3 を見てみよう．Y 字型の図形が描かれているが，この図形の中でどの部分に注目するだろうか．おそらく多くの人が，端の 3 点と線が分岐する部分に注目するのではないかと思う．これらの部分を注目する理由は，図形の特性が大きく変わる部分であるからである．前者は道の行き止まりになっており，後者は分かれ道になっている．位相幾何学の分野では，図形のこのような特性の変わる部分のことを臨界点と呼んでおり，その臨界点を見ることで図形の特徴を捉えようとした位相幾何学の分野がモース理論である．モース理論について誤解を恐れずに一言で言えば，図形の特徴を知りたければ臨界点を見ればよいということである．

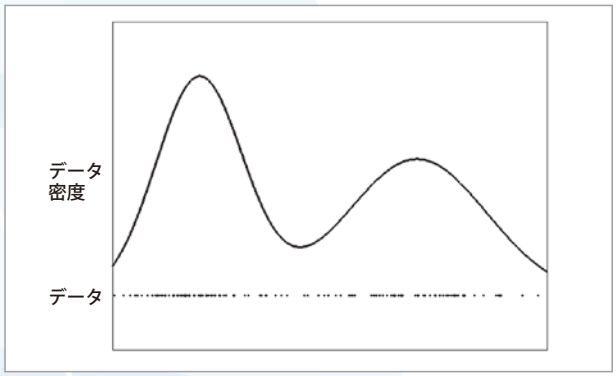

図 -6　下部に 1 次元のデータの分布，上部にはデータの分布の密度をプロットしている

さて，TDA ではデータの集合を図形として考えることでデータを解析しようとしていた．そこで，データという図形にモース理論の考え方を導入すると，「データの臨界点」を見ればデータの特徴が分かるということになる．この考え方をもとにした技術が TDA における Mapper 技術である．Mapper はデータの臨界点付近のデータをまとめて 1 つのノードとし，繋がっている（連続したデータのある）ノード間をエッジで繋ぐことで，データの集合をグラフに変換する技術である．ただし，臨界点付近にないデータも表現に残すために，エッジ上に臨界点以外のデータをまとめたノードを（1 つまたは複数）作成することもある．一般的にグラフは（エッジの重なりを許せば）2 次元に表示可能であるので，たとえ元のデータの次元が大きい場合でも Mapper により作成されたグラフは 2 次元で可視化が可能である．

具体的な例で見てみよう．**図 -4** では，Y 字型に分布したデータとなっている．このデータを図形としてみた場合の臨界点は 3 つの端点と分岐する点であり，Mapper は**図 -5** のように，それらの周辺をまとめたノードと，それらを繋ぐエッジで構成されたグラフを出力している．

Mapper を別の視点で考えてみよう．**図 -6** の下部のように 1 次元にデータが分布している状況を考える．ここで，図 -6 の上部のように，y 軸上方向に各データの周辺の密度をプロットしよう．これはデータ解析の上では，データ発生の確率密度関数を見ていることに対応する．このように，データの各点に対して密度などといった見たい情報を対応させたものをフィルタ関数と呼ぶ．図 -6 のようにフィルタ関数からできた図形に Mapper を適用すると，**図 -7** のようなグラフが出力され，

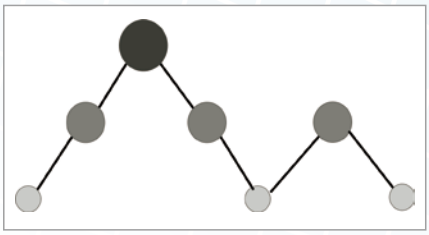

図 -7
図 -6 のデータに，密度をフィルタ関数とした場合の Mapper の出力例．ノードの濃さが密度の高さを表している

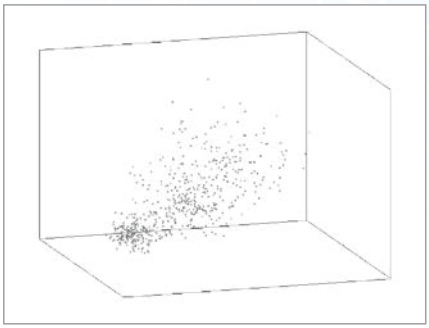

図 -8
3 次元空間上に 3 つの混合ガウス分布に従って発生させたデータ

各ノードには対応するフィルタ関数の値の範囲と属するデータの情報が保存されている．このグラフの各ノードに対応するフィルタ関数の値（ノードの濃さが対応）を見てみると，大小が交互になっており，フィルタ関数の山と谷の数を抽出することができる．

この結果はデータ発生の密な部分が 2 カ所あることを示しており，その結果全体のデータ発生の状況も捉えることができる．これは，従来の統計的な手法では確率密度関数を混合ガウスモデルによるガウス分布の数とそれぞれの平均を求めていることに対応する．しかし，混合ガウスモデルでは適切なガウス分布の数を決定するためには膨大な計算時間を必要とすることがあったり，複雑な確率密度関数の場合，ガウス混合モデルで表現することが難しい場合がある．一方で，Mapper を用いると具体的ではないものの，最も必要な情報を取り出すことができるのである．

この例では 1 次元のため，元々データをプロットすることで分かることではあるが，データが 3 次元以上になった場合にはデータの理解の大きな助けとなる．参考として，**図 -8** に 3 次元空間に分散の異なる 3 本のガウス分布を組み合わせた混合ガウス分布によって発生させたデータ，**図 -9** にそのデータの密度をフィルタ関数とした場合の Mapper の出力例を掲載する．繋がっているノードより密度の濃いノード（密度は色の濃さに対応している）が 3 つあるため，高さの違う 3 つの山があることが見てとれる．

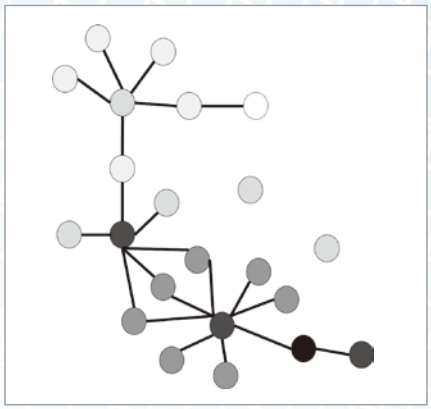

図 -9
図 -8 のデータに，密度をフィルタ関数とした場合の Mapper の出力例．ノードの色の濃さが濃度の高低に対応している

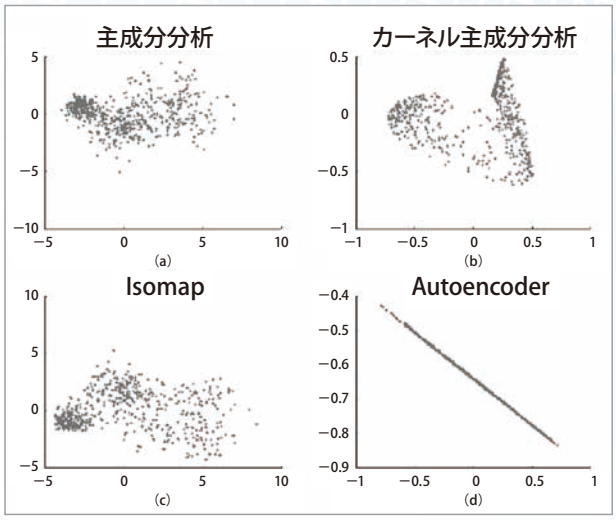

図 -10　図 -8 のデータの各種手法による次元圧縮結果

ビッグデータなどのデータ解析をする際，最もよく行われる解析の 1 つが次元圧縮による可視化であるが，先に述べたように Mapper 技術による出力は解析結果が容易に 2 次元で表示できるため，次元圧縮による可視化の技術として捉えることができる．

高次元データの次元圧縮の方法としては主成分分析が有名であり，さらに非線形な分布に対応したものとして主成分分析を非線形用に拡張したカーネル主成分分析，微分幾何的な概念を導入した Isomap などの多様体学習，ニューラルネットベースの Autoencoder などがある．これらはデータとの相性があり，必ずしもうまくいくとは限らない．**図 -10** は図 -8 のデータを上記 4 手法を用いて次元圧縮し，2 次元で表示したものである．いずれも 3 つのガウス分布から構成されていることを読みとることは難しいものとなっている．一方で，図 -9 では 3 つの山からできていることが分かるなど，より重要な情報を取り出すことができる．このように Mapper

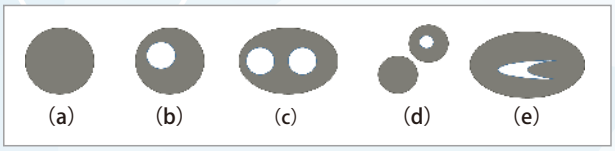

図 -11　穴の数の違う図形の例

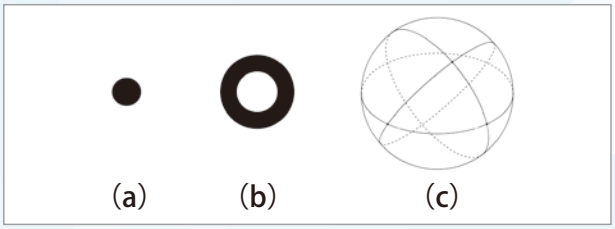

図 -12　穴の例．(a) は 0 次の穴を意味する連結成分，(b) が 1 次の穴を意味する一般的なイメージの穴，(c) が 2 次の穴を意味する球面

技術はデータサイエンティストなどがビッグデータから従来の手法では分からなかった有効な情報を取り出すツールとしてさまざまな可能性を秘めている技術である．

❖ データサイエンティストの道具として

Mapper 技術はデータサイエンティストがビッグデータなどに関するコンサルティングを行う際のツールとして，すでにさまざまなところで利用されている．特に TDA 技術の発祥の地であるアメリカではベンチャー企業を中心に適用事例が報告されている（文献 2）を参照）．

これらの企業では，従来の技術のデータ分析では発見することのできなかった癌の予兆や薬の適応の特徴，マルウェアによる攻撃の検知，金融ストレステストによるリスク管理への適用など，多くの分野で効果を上げ始めている．Mapper 技術はこれらの問題を直接解決するというものではないが，ビッグデータを解析するデータサイエンティストに，従来にはない有益な情報（insight などと呼ばれている）を提供してくれているのである．

データの穴が教えてくれること

❖ データ全体の形を捉える
—パーシステント・ホモロジー—

モース理論や Mapper 技術は，データ集合の中でどのデータを見ればよいのかを教えてくれるものであった．その一方で，データ集合全体としてどのような特徴を持つのか，異なるデータ集合がどのような違いを持つのか知りたい場合もある．そのような場合，どのようにしてデータを見ればよいか考えてみよう．

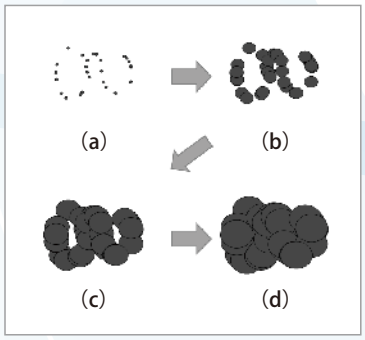

図 -13
パーシステント・ホモロジーのイメージ．球の直径が大きくなるにつれて (b) で中央に穴が生まれ，(c) ではすべてが連結し，中央の穴が消滅し左右に穴が生まれ，(d) ですべての穴が消滅し 1 つの塊となる

図 -11 を見てみよう．これらの違いを認識する際に，(a) ～ (c) の違いは穴の数で，(d) との違いは連結した部分の数と考える人は多いだろう．このように図形の形を「穴の数」を数えることで，図形の全体の形を捉えようとしたのがホモロジー理論である．ここでいう「穴」とは数学的には n 次球面と同相のものであるが，分かりやすくいえば**図 -12** のように 0 次の穴として連結成分の数，1 次の穴として一般的な穴としてイメージされるもの，2 次の穴として周りが密閉された空洞（球面）の数を数えたものである．

ホモロジー理論はこれらの数が同じであれば，大雑把な意味で同じ形をしているというものである．大雑把と言っているのは，図形がくっつきはしないが伸び縮み自由な素材でできていて，伸び縮みで変形したものは同じとしているためである．そのため，ドーナツと持ち手が 1 つのマグカップが同じとみなされていたり，図 -11 では (b) と (e) が同じものと捉えられる．

ホモロジー理論をデータ集合に応用すると，データ集合全体の形を大雑把に捉えることが可能になる．しかしながら，図 -11 の (b) と (e) を区別できないことはデータ解析をする上では都合が悪い．また，データの集合自体は点の集合にすぎないため，どのようなものを穴とするかも問題である．そこで**図 -13** のようにそれぞれの点を中心にボール状に膨らませていき，くっついた部分は一体化させてできた図形の穴の数（ホモロジー）を計算することを考える．膨らませるボールの直径に応じて穴の数は変化していき，穴ができてはつぶれていく．このように，ボールの直径に対応してできる穴の数の変化を見ることでデータの全体の形を捉えようというのがパーシステント・ホモロジーである．特に，ボールの直

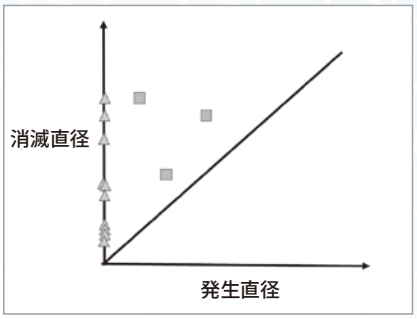

図 -14
パーシステント・ダイアグラム．横軸を穴が発生したときの球の直径，縦軸を穴が消滅したときの球の直径．0 次の穴を三角，1 次の穴を四角の点で表示している

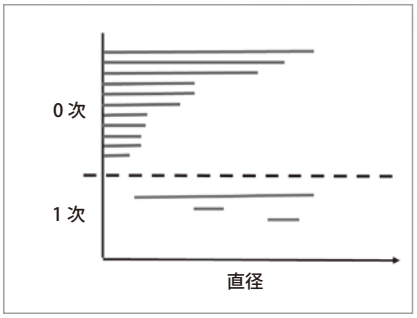

図 -15
バーコード．上段に 0 次の穴，下段に 1 次の穴を表示している

径に対する穴の発生と消滅の情報をデータ集合の特徴を表す情報として使おうというものである．これらの情報はデータの形を表現する“数値”であるため，機械学習などの手法に従来は取り入れることが難しかったデータの形に関する情報を加えることが可能になり，より高度な分類が可能になる，そのほかにも，データの形に起因する現象の分析もより高度化できる．

これらの情報は可視化しておくと便利であるので，代表的な可視化の方法として**図 -14** のパーシステント・ダイアグラムと**図 -15** のバーコードを紹介しておこう．パーシステント・ダイアグラムは横軸を穴の発生したときの球の直径，縦軸を穴の消滅したときの球の直径とした 2 次元座標平面上に，それぞれの穴の発生直径と消滅直径を座標上の点としてプロットしたものである．また，バーコードは横軸を球の直径として，各穴に対して発生直径と消滅直径を結んだものを並べたものである．

パーシステント・ホモロジーの重要な点は，膨らませたボールの半径に対する穴の数の「変化」を捉えたことで，単純なホモロジーのような大雑把な形を捉えるのではなく，具体的な点の配置の情報が分かるようになっていることである．たとえば，**図 -16** のように 3 つの点を正三角形と二等辺三角形上に並べたものを考えた場合，3 点を結んだ図形のホモロジーは 2 つとも同じものになるが，パーシステントホモロジーは 1 次

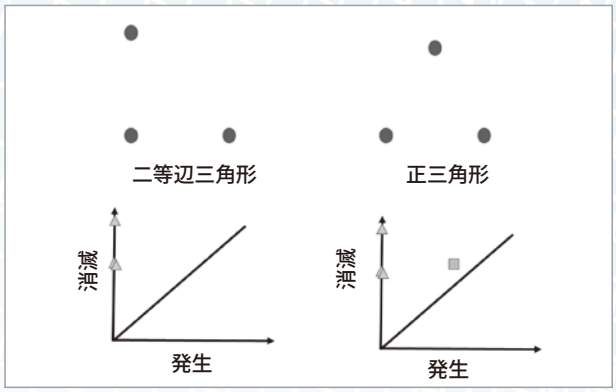

図 -16　3 点が二等辺三角形と正三角形に配置されたデータとそのパーシステント・ダイアグラム

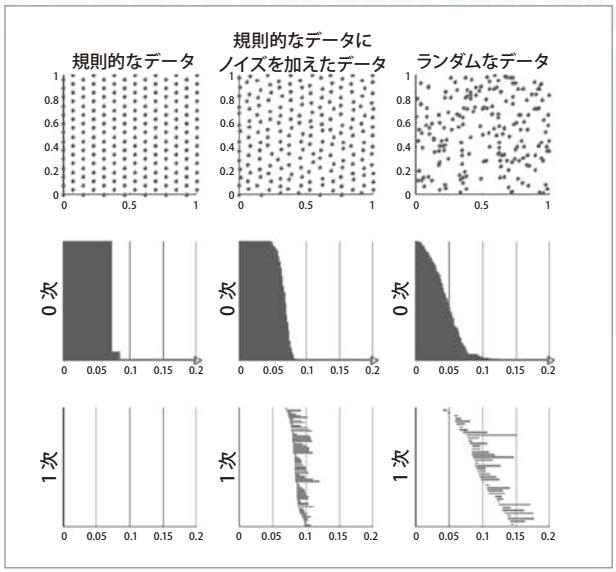

図 -17　規則的に並んだデータおよび規則的に並んだデータにノイズを加えたデータとランダムに並んだデータおよびそれぞれのパーシステント・ホモロジーのバーコード表示

の穴の発生の有無という異なる情報を与えてくれる．もちろん，多くのデータの中で少数のデータが少しずれた程度では，同じ結果を出力することがある．しかし，このようなデータのずれは一般的にノイズの影響と捉えることが普通であり，逆にノイズに対してロバストな技術となっていると考えられる．

パーシステント・ホモロジーはデータ集合の配置やバランスの情報を数値として取り出すことのできる技術である．データ解析への応用に関しては，データの集合がどのような配置やばらつき方をしているか知りたい場合に威力を発揮する．たとえば，**図 -17** のように規則的に並んだデータとランダムに配置されたデータの違いが，バーコード表示によってはっきりしてくる．以降の章では，これらの特徴を使ったパーシステント・

ホモロジーを応用した例について紹介する．

❖ 物質の構造を見極める

データの配置情報を捉えたい対象として，物質の構造は分かりやすい対象である．たとえば同じ物質であっても，分子の配置によって固体・液体・気体と変わってくるし，同じ個体であってもたとえば炭素であれば，通常の炭素とダイヤモンドの違いができたりする．現在，分子配置をパーシステント・ホモロジーを用いて解析することで，それぞれの物質の違いを解明しようとする動きが進んでいる（文献 3）を参照）．

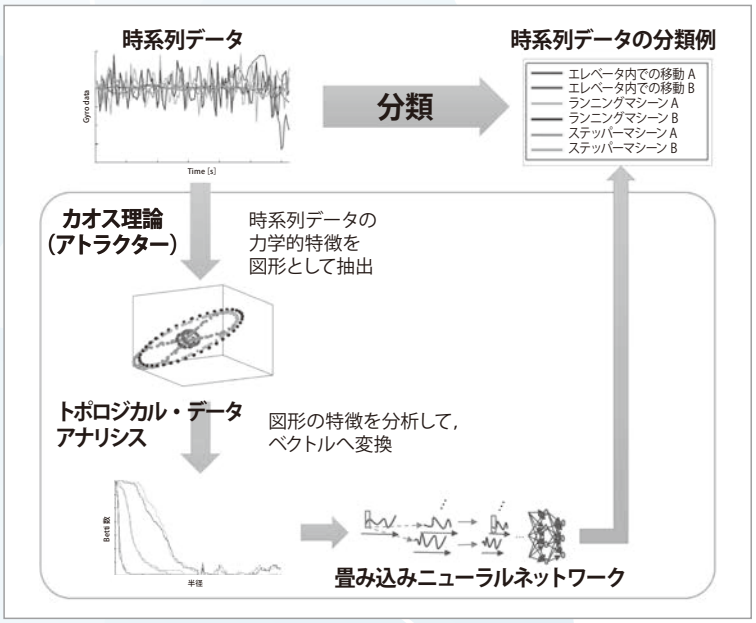

図 -18　パーシステント・ホモロジーを用いた時系列分析の例

❖ 時系列の特性を見極める

IoT 時代と言われる中で，数多くとれるデータがセンサのログデータなどの時系列データである．その中でも，激しく振動するようなセンサデータはカオス性（ルールには則っているものの一見不規則な動きをする）を持つものも多く，通常の解析手法では困難な部分があった．最近になり，時系列の発生のルールを図形化するアトラクタに対しパーシステント・ホモロジーを適用し，ディープラーニングと組み合わせることで高精度な時系列解析が可能になった（**図 -18**，文献 4）を参照）．パーシステント・ホモロジーの産業応用としては初めての例であり，今後医療・介護分野や製造分野・金融などへの応用が期待されている．

TDA の進む道

今まで見てきたように，TDA はデータ解析の道具として大きな可能性を秘めた技術であり，すでにさまざまな成果を出してきている．世界的にも学術界ではさまざまな分野で取り上げられ，ブームになりつつある．しかしながら，まだまだ未発達の分野であり，どのように使えばいいのか，何が分かるのか分からない部分も多い．そのため，TDA がより発展していくためには学術界・産業界両面でのさらなる研究・開発が必要である．しかし，現在の段階ですでに今までにない成果を出してきていることを考えると，研究・開発が進んだ先にはディープラーニングなどと並んで AI 社会の基盤となる技術となり得る可能性を秘めている．

本稿では，紙面の都合で各技術の数学的な説明や具体的な解析内容についての説明は避け，TDA で何ができるのか，そしてデータ解析にどのように使われるのかをイメージできるようになることを目的に TDA に関する導入部分の内容を解説した．より詳細な内容を知りたい方は，参考文献などの書籍や論文・報告書などを参照していただきたい．本稿を通じて興味を持ち，自らの研究や業務に取り入れてくれる人が増えれば幸いである．

参考文献

1) Carlsson, G. ：Topology and Data, PBULLETIN OF THE AMERICAN MATHEMATICAL SOCIETY, Vol.46, No.2, pp.255-308 (Apr. 2009).
2) Nielson, J. L. et.al. : Topological Data Analysis for Discovery in Preclinical Spinalcord Injury and Traumatic Brain Injury, NATURE COMMUNICATIONS, DOI:10.1038/ncomms9581 (Oct. 2015).
3) 平岡裕章：タンパク質構造とトポロジー―パーシステントホモロジー群入門―，共立出版 (2013).
4) 富士通（株）：人々の安心安全な暮らしを支える新しい AI「時系列ディープラーニング」, FUJITSU JOURNAL (Mar. 2016).

（2016 年 7 月 29 日受付）

梅田裕平　umeda.yuhei@jp.fujitsu.com

2009 年九州大学博士課程修了．2010 年より（株）富士通研究所研究員．現在は人工知能関連の研究に従事．