文化計量学実習での分類と系統の観点

Phylogenetic point of view for cultural metrics

榛葉　豊*

Yutaka SHINBA

Abstract : We argue the theorem of ugly duck by S. Watanabe to think about the affinity, so that the multivariate analyses, especially Hayashi's quantification methods, are discussed. The usage of the Phylogenetic point of view in the cultural metrics is analyzed.

1．多変量解析，特に数量化理論

静岡理工科大学出は人間情報デザイン学科のカリキュラムに多変量解析関係や科学計量学，世論分析，マーケッティング関係などの科目が複数用意されている．その中で筆者は，人間情報デザイン実験を担当している．実験では文化計量学的題材の中から，統計的文体論であるとかテキストの情報理論的分析を実施している[1-8]．

学科の卒業研究でも人間情報デザイン学科では，心理学，感性評価，テキストマイニング，社会調査，教育学など，人間の行為に関する統計的手法を含むテーマを実施している研究室が多い．データの山からコンピュータ上のツールを用いて，対象を何らかの観点で2次元平面などに布置するということは，現在では手軽にできる．複数の観点から対象集団を計量し，（特に主成分分析，因子分析などでは）その多変量データから対象の特徴を先ずは1つの変数に集約するのである．次の段階ではさらに残された情報から第2の合成変量を作って，第1と第2の変量のなす平面に対象を，そして元の変量達も，布置して分類考察するのである．

多変量解析の手法には具体的には重回帰分析，主成分分析，因子分析，判別分析，クラスター分析などがある．（多変量データが，実数で表されるような変数では無く，yesとnoのような質的データである場合には数量化理論といわれる手法が用いられる．）重回帰分析と判別分析は，外部から与えられた目的変数があり，それにたいしどの説明変数がどのように効いているのかという事，すなわち因果関係の推定が問題になる．それに対し主成分分析やクラスター分析では，外部から与えられた目的変数は無く，対象集団についての計量されたデータ自身から分類を行うのである．

しかしそのようなことをコンピュータで実行させていると，手軽すぎて，どういう原理を用いて，何のために，何をどういう空間に布置しているのか，ということはあまり学生の関心の対象として浮かんでこないようである．ここでいっているのは勿論数学的な手法の詳細などという事では無く何をやっているのかの理解ということを言っているのである．

学問の基本は，そして実務の基本も同じであろうが，標本採集，未知の現象の発見と観察などの博物学的段階をすぎたら，とにかく分類するということである．その後に理論構築が来て，仮説演繹法の最終段階の実証が来るのである．

分類というのは，何かの意味で「似ている」対象を分別することである．そこで，似ているというのはどういう事なのかが，大切な問題になるのである．

2．類似しているとは

近代西欧科学の一つの特徴は，対象や現象を数量化して記述するということである．ある対象集団を，ある観点で数量化して類似性でグルーピングしようという場合，その特定の観点から漏れたそれ以外の特徴は捨てているのである．そうでなければ研究の第一歩が始まらないわけである．そうした態度こそが，要素還元論的態度の特徴である「第一近似」という考え方である．

こうしたわけで数量化されて，各個体に与えられたた数値（複数の観点を採用しているのならベクトル）を根拠

2017年3月7日受理

*　総合情報学部　人間情報デザイン学科

にして，似ているとか似ていないとかを判断することになる．その際に取られる判断法は「距離」という概念によっている．（数学的には，距離よりもっと一般の，位相という概念になる．）2つの個体間の距離とは，その2つの個体が似ていない度合いを数量で表すもので，逆にいうと，距離が小さいのなら似ているということである．その2つの個体xとyとの距離をdと書くことにする．この数値は負の距離ということはないのだから正の実数で無ければならない．最も似ている，ということは距離が0と言うことで，似ていない度合いが増すごとにその数値は増えていくのである．　すなわち$d(x,y)\geq 0$　である．この汎関数に対する要請は，　距離の3公理といい，次の3つである．

$$d(x,y)=d(y,x)$$
$$d(x,y)=0 \quad \Leftrightarrow \quad x=y$$
$$d(x,y)+d(y,z)\geq d(x,z)$$

一番目は，2つの個体のどちらを基準にしても似ている度合いは同じであるということ．第2のものは距離が0ならそれらは同じもの（と見なす）．また逆に同じものの似ていない度合いは0である．3つめは三角不等式といい，第3の個体を考えたとき，第1のものと第2のものを直接考察したときより，第3のものを導入してそれ経由で考えたときには，直接考えたときと同じかそれ以上であるということである．平面上の距離なら自明なことである．

このような距離という，似ていない度合いの指標を用いて，クラスター分析ではグルーピングを行うのであるし，また他の多変量解析の手法でも個体間の距離という概念は頻用される大切なものである．

よく使われる距離には，平面上の距離と同じユークリッド距離も勿論のこと，市街地距離などいろいろな形の距離が用いられる．また距離の3公理のうち，第3のものだけ満たさない擬距離というものを採用することもある．統計的文体論や科学計量学[9]では，距離として共出現（共起）という概念が使用されることも多い．例えば，科学計量学では2つの単語それぞれが出現する論文の数をC_i, C_jとし，両方の単語が出現する論文の数をC_{ij}とすると，

$$J_{ij}=\frac{C_{ij}}{C_i+C_j-C_{ij}} \quad , \quad \text{ジャカール指標}$$

であるとか，

$$S_{ij}=\frac{C_{ij}}{\sqrt{C_iC_j}} \quad , \quad \text{コサイン指標}$$

のような擬距離指標で，ある研究分野の論文集号の中での2つの単語の近縁度を表したりする．1つの小説の中での2つの単語の近縁度というのなら，論文数が小説中の文の数ということになる．この近縁度を使って，単語の地図を作成したりする．論文の共著者ということを用いれば，同様にして学者の共著地図が作成できる．

符号化理論では誤り訂正などでハミング距離というものが用いられる．ハミング距離は，2つの同じ桁数の文字列間で，異なった文字が用いられている桁の数である．数量化理論のように質的変数，特にyesとnoの2値の場合（0と1という2進数一桁で表される）だと，これは2つの2進数の間で，0，1が異なっている桁の数である．

このように，色々な分野で類似度を測る指標としてさまざまな距離が用いられているが，いずれにせよ，2つの個体について測定した複数の変量から，全順序である（任意の2つの個体は必ず比較ができて，順序づけができる）ところの，距離なる1つの数値を使って類似度の大小判断をするというわけである．これが分類の大本である．

3.醜いアヒルの仔定理

ところがここに，渡辺慧によって証明が与えられた「醜いアヒルの仔定理」[11-13]というものがある．この定理は，個物はどの2つをとっても同じぐらい似ているという驚くべきことを主張している．（醜いアヒルの仔は実は白鳥の子供なのであるが，仔アヒルの中に混じっていると区別がつかない．アヒルの仔同士の類似度と，アヒルの仔と白鳥の仔の類似度は同じであるということからの命名）簡単な場合で説明しよう．

まず，より似ているとは，共通に当てはまる形容詞の数が，より多いことであると定義する．

さて，4つの個物を分類することにする．そのためには最低2つの形容詞（述語）がなければならない．それをA，Bとする．個物がその形容詞に当てはまるかどうかで，0または1の値を取るとすれば，2つの形容詞の組み合わせで4つの個物を識別できる．

次にAとBから作られる原子命題を考えよう．原子命題とは，それ以上他の命題の選言（「または」）として作られない命題である．具体的に言うとベン図上で区別された領域ただ1つに対応する命題のことである。形容詞が2つの場合これは4つある．すなわち　$\neg(A\vee B)$，$A\wedge\neg B$，　$\neg A\wedge B$，　$A\wedge B$である．これらの4つの原子命題のうち（異なる）2つから作られる選言（2位の述語という）は6つある．それは　A，　B，　$\neg A$，　$\neg B$，$(A\wedge\neg B)\vee(\neg A\wedge B)$，　$\neg((A\wedge\neg B)\vee(\neg A\wedge B))$　である．

最後から2つめは排他的論理和である．原子命題3つの選言（3位の述語）は4つで，　$A\vee B$，　$\neg A\vee\neg B$，　$A\vee\neg B$，$\neg A\vee B$となる．以上14個に，恒真命題 □ と恒偽命題 ◎ を加えた16個の命題はブール束を作っている．これ

ら16個の命題を拡張された形容詞という．

この拡張された形容詞群は，分類が分類者の持つ嗜好やものの見方に影響されないためには，重要度の点において対等に扱わねばならない．

さてここで4つの個物の中からどの2つをとっても，その2つの個物に共通に当てはまる形容詞の数，すなわち1の数は同じで，16個の拡張された形容詞中の4つであることが表を書いてみればわかる．16個の拡張された形容詞に当てはまるか当てはまらないかを順に記して，個物を16桁の2進数として表せるということである．ということは4つの個物の中から任意に選んだ2つの個物間のハミング距離は12である．

4つの個物は最初の仮定から，2つの形容詞から作られた原子命題の1つ1つに対応している．2つの個物に共通な原子命題は定義からあり得ない．拡張された形容詞のうち2位の述語には，まさに2つの個物が対応するそれぞれの原子命題の選言1つが共通である．3位の述語ではその2位の共通述語にもう一つ別の原子命題を加えたものである．これが2つある．最後に恒真命題を付け加えて全部で4つの命題が2つの個物に共通に真となるというわけである．

白鳥の仔1羽とアヒルの仔3羽がいるとすると，アヒルの仔どうしも区別できるのならば，アヒルの仔同士でも白鳥とアヒルでも類似度はまったく同じということを主張しているのである．このことは，個物数が多くなっても同様に成り立つ．つまり，分類という行為は，それが客観的あるいは先験的なものであるのならば，不可能であるということである．

勿論現実には何らかの意味で「客観的」な，距離による分類は可能である．形容詞，すなわち変数の間に重要度の差があれば醜いアヒルの仔定理は成り立たない．我々が行う分類という行為は，変数は限りなく沢山あるのにその中から，何らかの変数を選び出して計測して分析するということなのであって，それ自体が恣意的であったり，もしくは分析者の価値観を強烈に反映した行為なのである．

多変量解析をするとき，説明変数として何を取るかということはどのように決定されるのだろうか．先行研究でそれらが採用されていたから，それらが計測しやすい変数だから，なんとなく思いついた変数，その研究対象について頭の中で考えて，それらの変数にこそ違いが出るであろうと思われる，等々．ということもあるであろう．だが，変数の取捨選択ということこそが（変数の間の相関などによる，数学的な計算上の問題は別として），分類を決めているのだということを肝に銘じなくてはならない．研究で大切なのは，先ずは変数を熟慮の上で決めて，多変量解析なりをして，重要な変数，無駄な変数を探る．そして可能なら元に戻って新しい変数を見つけ，それを計測する手段を開発するというサイクルを行うことなのである．

4. 数量化理論Ⅱ類，Ⅲ類では何を数量化しているのか

数量化理論Ⅱ類は，数量的変数の場合の線形判別分析に対応するもので，2値の複数の変数から1次の判別関数に相当する合成変数を作りその値によって（多くの場合，正負の別でという形にする）2つのグループに分ける．この合成変数を作る原理は次の通り．

外部から与えられた2値の目的変数の値で分けた2グループが，その合成変数値の上でできるだけ離れるようにする．この分離具合は相関比ηと呼ばれる量で表される．つまりηが最大になるように合成変数（サンプルスコアと呼ばれる）の係数を決めるのである．

結果として得られたサンプルスコアという数値はグループの個体の持つ特徴の強さを表しているのだから，サンプルスコアの差は距離を表している．この距離を使って，個体についてのさらなる議論もできる．

サンプルスコアを与える関数は，N個の説明変数の分析の時，個体を表すN桁の2進数に対してサンプルスコアという実数値を与えるものだと言ってよい．

一方，数量化Ⅲ類は，数量的変数の場合の主成分分析に相当し，分析者が外部から与えられた目的変数は無い．個体と変数を縦横の並びとする表で，対角線に近いところに反応が集まるように，個体と変数の両方を並べ替え，相関が高くなるようにすることを目標とする．そのやり方は，個体の数量化ウェイト，変数（カテゴリー）のウェイトを考え，その間の相関が最大になるようにウェイトを決めるのである．この問題を線形代数で扱うと複数の解（固有値）が得られるが，その最大のものを1軸，次に大きなものを2軸と呼び，1軸—2軸平面に変数達，または個体達を布置して考察するのである．すなわち平面上で固まっている個体は似た個体のグループであり，同じ平面に変数を布置した図からは，平面上で固まっている変数のグループは，多くの個体に対して同じような値を与える，似た変数なのである．

Ⅲ類の場合も，Ⅱ類同様であるが，1軸，2軸など．あるいはこの平面上のユークリッド距離は，距離が近いもの同士は似ているという分類をするための距離として機能しているのである．個体を平面に布置しているときにはやはり，個体を表す2進数に対して，数量化した1軸の値，2軸の値をきめ，平面上の位置を与えているのである．

5.分類学と系統学

以上の節で，多変量解析や数量化理論で個物を分類するという行為について何を強調して教えるべきか考えてきた．しかしそこで言及したことはすべて，「今」という時間軸上の切断面上での，現時点での類似からの分類である．そこには歴史とか由来とか言う系統学の観点は全くない．生物の分類で，現在観察されている形態からの分類というのと同じである．だが，もし化石というものが発見されたらどうであろうか．その場合には過去の事物との整合性が

とれるように分類をしなくてはならない．また現時点でのデータだけから分類する場合であっても，ある生物種はずっと昔から変化していなくて，あるものは途中から分岐したものなのかも知れない，そうならある種は別の種と現時点で類似しているというだけではなく，その先祖なのかも知れない，などといういろいろなことが出てくる．

特に生物学でよく言われる，「相似と相同」という問題がある．例えば袋オオカミとオオカミは形態も生態学的地位もよく似た種であるが，片方は有袋類，片方はほ乳類と大きく異なる．これは相似である．器官についても，鳥の羽と昆虫の羽のように相似器官がある．逆に，現状が異なっても由来や先祖を同じくするものは相同である．人間の手とコウモリの翼は相同器官である．このような事を考慮すると，歴史を復元するという系統の推定は非常に難しい．

遺伝子の系統というようなことになると，系統推定法はいろいろなアルゴリズムが開発されて使用されているが，考え方に大きな違いがあり，分岐分類学，進化分類学などいろいろ考え方がある．樹形図もそれが系統を表しているのか単なる分岐図なのか，意味はまったく異なる．

我々は卒業研究などで，典型的には，写本の系統分析のような，文化計量学分野のことを行わせようとするのであるから，この分野の手法を取り入れるだけでは無く，上記のいろいろな事情や概念の理解を深めさせなければならない．系統学の方が優れていると言いたいわけではない．現時点での分類がすべてであって，人間にとって知りうるのはそれがすべてなのだという立場も勿論あり得る．現時点で見いだし得ない差異しか無いのなら，「本当は」詳細な差異があるというのかもしれないいが，それらは現時点では同じなのかも知れない．

系統ということについては，いろいろな立場があろうが，本学の実習では系統の観点を含んだ実習を開発していきたい．

参考文献

1） 榛葉豊，『分類すると言うこと』静岡理工科大学紀要，第14巻，(2006年)
2） 榛葉豊，『テクスト間のクラスター解析における Kullback - Leibler Divergence 型距離』，静岡理工科大学紀要，10巻（2002年）175-185
3） 金明哲，『テキストデータの統計科学入門』，岩波書店（2009年）
4） 村上政勝，『シェークスピアは誰ですか』，文藝春秋（2005年）
5） 堀江，奥田，川村，山下，尾畑，「作者識別の一例」，『人文科学における数量的分析』，総合研究大学院講演録（1999年）
6） 小田晋，『グリコ森永事件 21 世紀型犯罪を分析する』，朝日出版社（1985年）
7） 前川守，『1000万人のコンピュータ科学：文章編』，岩波書店（1995年）
8） 山内保典，「科学事件により新聞報道はどう変わるのか？ ー考古学報道を題材にー」，科学技術社会論学会2005年大会（名古屋大学）予稿集（2005年）
9） 藤垣裕子他編，『研究評価・科学論のための 科学計量学入門』，丸善（2004年）
10） 増田直紀，今野紀雄，『複雑ネットワークの科学』，産業図書，(2005年)
11） S. Watanabe, '*Knowing and Guessing*', Wiley(1969) 村上陽一郎訳，『知識と推測－科学的認識論』，東京図書（1975年)
12） 渡辺慧，『認識とパタン』，岩波書店（1978年）
13） 渡辺慧，『知るということ－認識学序説』東京大学出版会（1986年）
14） 池田清彦，『分類という思想』，新潮社（1992年）
15） 村上政勝，『文化を量る』，朝倉出版（2003年）
16） 村上政勝，『真贋の科学』．朝倉出版（1995年）
17） E.ワイリー他，『系統分類学入門：分岐理論の基礎と応用』，文一総合出版（1992年）
18） 中尾央，三中信宏編著，『文化系統学への招待』，勁草書房（2012年）
19） E.ソーバー，『過去を復元する』，勁草書房（2010年）
20） 三中信宏，『系統樹思考の世界』，講談社（2006年）
21） 三中信宏，『分類思考の世界』，講談社（2009年）
22） 甘利俊一，長岡浩司，『情報幾何の方法』，岩波書店（1994年）