スパースモデリングの脳神経活動計測への応用

東京理科大学 理工学部 教養 講師　市川（いちかわ）　寛子（ひろこ）

はじめに

スパースモデリングとは，用いる変数をスパース（sparse，疎）にしてモデリングを行うことである。筆者が公募班員として参加する新学術領域研究「スパースモデリングの深化と高次元データ駆動科学の創成」のホームページ（sparse-modeling.jp/）には，スパースモデリングについて「その基本的な考え方は，(1) 高次元データの説明変数が次元数よりも少ない（スパース（疎）である）と仮定し，(2) 説明変数の個数がなるべく小さくなることと，データへの適合とを同時に要請することにより，(3) 人手に頼らない自動的な説明変数の選択を可能にする枠組みである」と説明されている。

より少ない変数で記述することにどのようなメリットがあるのか。果たして変数を削ることは本当に有用なのか。本稿では，具体的な例から考えてみたい。

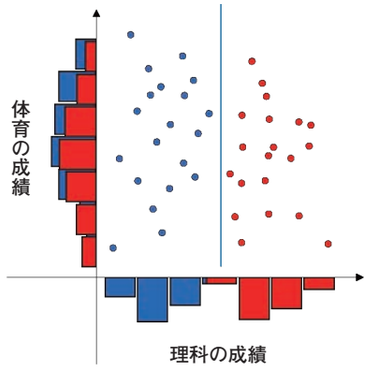

図１　理科の成績と体育の成績の散布図
各点は個人を表す。赤は理系，青は文系。

変数をスパースにすることの意味

「理系」は計算や理論的な思考が得意で，「文系」は語彙が豊富で情緒がこまやか，というイメージは多くの人が抱いている。「文系」「理系」という区別は，その人の特徴を表すことに役立つと考えられている（ちなみに筆者が専門としている心理学は，日本の大学では「文系」学部に属し，欧米の大学では「理系」学部に属する）。

ある人が理系か文系かを，学校で習う教科の成績から予想することを考えよう。**図１**を見ていただきたい。あるクラス40人の，理科の成績と体育の成績を散布図としてあらわした仮想データである。

理科の成績を見て，理科の成績が高い人を理系，低い人を文系と判断すれば間違いなく当てられることが分かる。体育の成績は，理系か文系かには関係のない能力であるから，識別のための変数として使わなくともよいようだ。図１の周辺分布をみても，理科の成績は分布の平均値がきれいに分かれており，理系と文系とをきれいに識別できることが分かる。体育の成績は，周辺分布の大部分が理系と文系とで重なり合っている。

これは，理科の成績，体育の成績の２つを変数として，理系／文系というデータのラベルを言い当てる問題である。このような問題を識別問題と呼ぶ。識別問題を解くためには，識別に必要な変数を選択する必要がある。今回は理科の成績が変数として選ばれ，体育の成績は間引かれた。

「いやいや，せっかく体育の成績も手元にあるのだから，情報を捨てるのはもったいない」と考えて，体育の成績も“無理やり”用いて識別しようとすると，**図２**に示すような識別面が作れる。ただしこの時，図２の左，右，いずれの識別面でも許される，つまり，識別面が一意に定まらない。

一意に定まらないなら，どちらの識別面を選んでもいいじゃないか，ということにはならない。よい識別とは，すでに取得したデータだけを正しく識別すればよいのではなく，これから取得される新規データもうまく識別できることが求められる。そこで，図２のそれぞれに，新規データを追加したのが**図３**である。

新規データが理系と文系のどちらのラベルをもつかは，識別面によって変わってくることが分かる。このように識別面が一意に決まらないことは，選ばれた変数からは，新規データに対する識別能力（これを汎化（はんか）能力と呼ぶ）が高い識別面を作れないことを意味する。つまり，体育の成績を変数として採用すると，識別面を決めるべき変数にノイズをまぜることになり，識別面の汎化能力を下げることになる。したがって，体育の成績は，せっかく計測できていても削るべき変数であると判断できる。

このように，問題を解くために必要な説明変数を選ぶ，すなわち不要な変数を削ること，これがスパースモデリングの基本姿勢である。

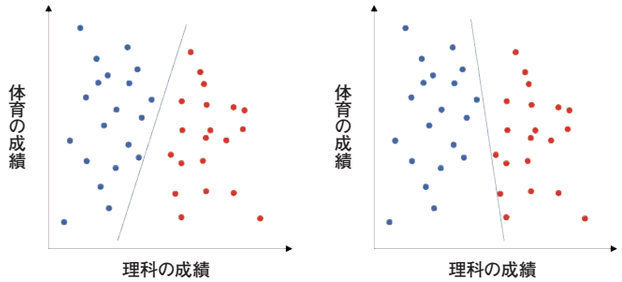

図２　理科の成績と体育の成績の散布図
直線は，理科の成績と体育の成績をもとにひいた識別面。各点は個人を表す。赤は理系，青は文系。

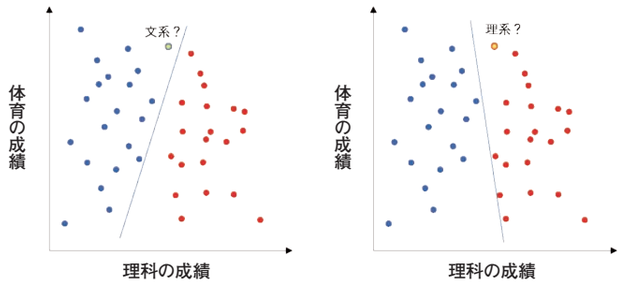

図３　理科の成績と体育の成績の散布図
新規データがどちらのラベルをもつかは，用いる識別面によって異なる。

変数は組み合わせをみながら選択するべきである

今度は，算数の成績と社会の成績から，理系と文系を識別する問題を考える。**図４**を見ていただきたい。先ほどの図１の凡例を引きつぎ，理系を赤，文系を青で示す。

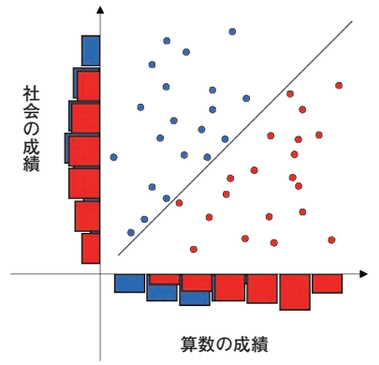

図４　算数の成績と社会の成績の散布図
各点は個人を表す。赤は理系，青は文系。

今回は算数の成績と社会の成績という２つの変数を使って作った識別面でうまく識別できることが分かる。

ここで興味深いのは，算数の成績と社会の成績，それぞれの周辺分布では，きれいに文

系と理系が分かれていないことである。これは算数の成績だけ，あるいは，社会の成績だけで，理系か文系かを識別することが難しいことを意味する。では，前節の体育の成績のように，理系か文系かの識別に寄与しない変数として，これらの変数を間引いたほうがよいのだろうか。答えは否である。

実は今回の問題では，１変数単独で識別力が低くても，他の２つを組み合わせることで，識別問題をうまく解くことができる。それは算数と社会の成績に相関があるためである。相関のある変数を組み合わせた合成変数は，個々の変数の識別能力よりはるかに高い識別能力をもつことがある。したがって，２つ以上変数がある場合には，変数単独での識別能力が低くてもすぐに間引くべきではなく，それらを組み合わせた合成変数の識別能力を検討し，十分に高ければ変数を２つそろえてモデルに入れることが得策である。

以下に，変数を組み合わせた場合の識別能力を検討する手法について紹介する。４つの説明変数をもとにデータを２群に識別する問題を想定し，五十嵐康彦ら（2015年，ヒューマンインタフェース学会誌）は人工データを作成し解析した。３つの変数はガウス乱数によってノイズを加えた“真の”変数で，残りの１変数はラベルによらずランダムに値を発生させたノイズである。**表１**は，変数を組み合わせた場合それぞれについての識別率を表したものである。

表１　用いた成績と２群の識別成功率

用いた変数	訓練データの識別率	新規データの識別率
A，B	82.5%	75.0%
A，B，C，D	100.0%	87.5%
A，B，C	100.0%	95.0%

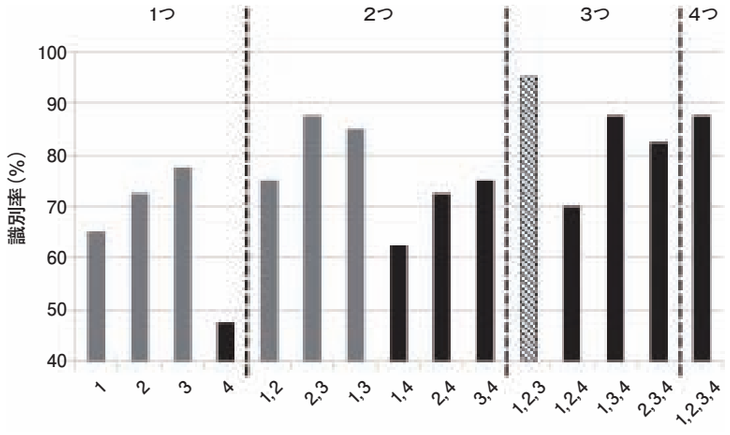

図５　４変数すべての組み合わせにおける新規データの識別率

変数A，Bだけの場合に比べて，すべての変数を用いた場合，あるいはA，B，Cを用いた場合に，訓練データの識別率は向上した。訓練データに対する識別能力は，変数を増やすほど向上するのが一般的である。一方で，テストデータに対する識別率は，すべての変数を用いた場合よりも，変数Dを間引いた場合のほうが高かった。これは，変数Dが識別に有用な情報をもたない変数であるため，図３に示したように，訓練データの選び方によって大きく識別面を揺らがせる原因となることで，新規データの識別率を低下させたことを示している。

識別面の決定に必要な汎化性能を検討するため，４変数のすべての組み合わせ（$2^4-1=15$通り）における，新規データの識別率を**図５**に示す。この図から，識別に必要な情報を持たない変数D（図中の変数４）を入れると，入れる前の組み合わせよりも識別率を悪化させることが，すべての変数組み合わせにおいて観察された。

全状態探索法

本節では，変数の組み合わせを考慮しながら厳密にスパースモデリングを行う唯一の方法といえる，全状態探索法と，これを用いた脳神経活動計測の解析例を紹介する。

全状態探索法とは，変数がD個のとき，識別面を作るためにつかう変数のすべて組み合わせ，すなわち2^D-1（$={}_nC_1+{}_nC_2+{}_nC_3+\cdots{}_nC_{n-1}+{}_nC_n$）通りの識別面を作り，汎化性能を評価する手法である。手法の最も困難な点は，変数が１つ増えるごとに計算量が２のべき乗で増えることで，膨大な計算量が必要になることである。

膨大な計算量を抑えるために，近似アルゴリズムが開発されているが，厳密な解をもたらさないことが指摘されている。また，よく知られている変数選択手法としてステップワイズ法があるが，識別力の高い変数から１つずつを取捨選択していく方法であるため，前節で紹介したような“単独では識別能力が低いが，組み合わせることで識別能力が上がる”変数を拾い出すという発想はない。その点，全状態探索では変数１つ１つが持つ，ほかの変数との相関関係を検討しながら，厳密に最適なモデルを選ぶことができる。

筆者らは，脳神経活動のデータに，この全状態探索を適用した。

Eifuku et al.（2004, *J. Neurophysiol*）は，マカクザルが４人のヒトの顔を観察しているときの側頭葉の一部の活動を，シングルユニットレコーディングにより計測した。

我々はこのデータを用いてある２人の顔を識別するときに活動するニューロンが23個のうちのどれかを見つけることを目的としてデータ解析を行った。すなわち，説明変数が23個で，人物Aと人物Bを識別するという識別問題である。これを近似アルゴリズムによって解いた場合には新規データの識別率は93%であったが，全状態探索では識別率100%となる変数組み合わせを複数見つけることができた。

全状態探索は，脳科学以外の分野，例えば，地質学にも適用されている（Kuwatani et al., 2014, Scientific Reports誌）。彼らは，2011年の東日本大震災の津波堆積物に含まれる元素含有量を訓練データとし，新規データが津波堆積物かどうかを識別する問題を全状態探索によって解いた。

人工知能（AI）でも全状態探索と同様のサンプリングアプローチを用いて問題解決を行っている例がある。最近であれば，2016年８月，IBM社の人工知能Watsonは，DNAマイクロアレイデータを用いて，遺伝子情報から患者の白血病のタイプを識別したことが報告された。その結果，患者の治療方法の変更を提案し，患者の回復に貢献したことが大きなニュースとなった（「AI，がん治療法助言　白血病のタイプ見抜く」日本経済新聞2016年８月６日版）。

まとめ

スパースモデリングの考え方，識別問題におけるスパース変数選択について紹介したのち，脳神経活動計測データの解析を行った例を紹介した。さらに，本稿のタイトルとはそれるが，スパースモデリングが，脳神経活動計測以外のすべての科学に適用可能であることも紹介した。

問題の型が同じであれば，その問題を解くための手法は研究領域を超えて適用できる。心理学の専門用語に，知識の領域固有性という言葉がある。問題解決に関する知識が学問分野などの個々の領域に固有のものと考えられがちで，ほかの分野で同じ型の問題があっても，知識を一般化して適用することが難しいという心の癖を指す。

しかしながら，宇宙のデータの解析と遺伝子データの解析など，天文学と生物学という異なる分野にも，同じスパースモデリングの手法が適用されているケースを，冒頭に紹介した新学術領域「スパースモデリング」では目の当たりにすることができる。今後は学内の多くの研究分野においても，もちろん著者が専門としている心理学においても，スパースモデリングが活用されることを期待している。