センターとデータ解析と私

岡田　孝（関西学院大学　理工学部情報科学科）

1. はじめに

著者が情報処理研究センターに奉職したのは，１９７６年であった．その後情報メディア教育センターに変わったが，来年度は，また大きく組織変えが行われるそうである．この情報科学研究も本号で終わりになるかもしれないと聞く．四半世紀以上もお世話になったセンターについては，思い出はつきない．現在の筆者は，データマイニングを講義しているが，赴任当時はデータ解析の言葉自体も知らない状況であった．データ関係についてはセンターで育てて頂いたと申して過言ではない．そこで，センターとデータ解析について筆者の知るところを記して，センターへのささやかな感謝の思いとしたい．

2. センター草創期のSPSS

1976 年は，情報処理研究センターが設立され，また全学共用の中型汎用ホストコンピュータ Facom230-38 が導入された年である．筆者は当時，プログラムは自作か，研究仲間で共用して使うのが当たり前と考えていた．ところが，このコンピュータの利用者の多く，特に文系の先生方は，センターの雄山先生が作成された KGSL（Kwansei Gakuin Statistical Library）という共用のプログラムを利用されていた．ユーザの多いプログラムがあるものだと感じていたが，その内容は全く知らなかった．

驚いたのは半年か１年ほど経って SPSS が利用可能になった時である．現在のような商用パッケージではなく，当時はシカゴ大学が開発したオープンソースのソフトウェアであったと思う．SPSS を日本でも使用可能にすべく，当時京都大学の三宅先生が中心となって科研費プロジェクトを立ち上げられ，京都大学の富士通製大型機用にコンバートされていた．雄山先生もこのプロジェクトに参画され，富士通製中型機へのコンバージョンが関学のセンターで行われていた．学外からの山本先生や安田先生のご助力もあって，関学では全国大学のトップを切って利用が可能になったのである．

このソフトウェアが動き出したとたんに，これまでの KGSL 利用者が雪崩を打って SPSS に移行し，利用者数も計算ジョブの数も急増したように思う．このように多数の利用者がいるソフトウェアとは一体何をするものかと疑問を持ち，雄山先生に「多変量解析ですよ」と教えていただいたのが，思い起こせば筆者とこの分野とのつきあいの始まりであった．

標準的な教科書である奥野忠一先生の「多変量解析」を読み，改めてびっくりした．筆者は院生時代に量子化学分野で分子軌道法の勉強をしたのだが，多変量解析で使われている方法論と分子軌道法のそれが見事に対応していたのである．主成分分析は分子軌道法そのものに，また因子分析は局在化分子軌道法，正準相関分析は対応軌道法というように，ものの見事に構成が一致していた．線形空間の問題にしてしまえば，異なる分野であっても同じということを初めて認識したのは，恥をさらすようではあるが新鮮な体験であった．論文として提案された年代を見てみると，数年程度の差ではあってもほぼ多変量解析の方が早く，分野が違えば方法を導入するだけでも一流の仕事になるということを感じたものである．同じことは今でも続いており，例えば主成

分分析がパターン認識では部分空間法としてもてはやされ，バイオインフォマティックスでも華々しく再登場したのは比較的に最近のことに属する．

SPSS に関しては，林知己夫先生の数量化理論がソフトウェアでカバーされていないことが問題になっていたが，後でその機能が追加されたように思う．併せて，数量化理論とコレスポンデンス分析や dual scaling との関係もお聞きし興味深かった．

3. センター充実期とSAS

1980 年代に入って，センターのホスト計算機も M シリーズの IBM 互換機に変わり，世界中のソフトウェアが自由に使えるようになった. 正確な年代は記憶にないが, また雄山先生が「SAS をセンターで導入すべきである」と主張された．当時 TSS 処理が可能になり，SAS Graph で結果をグラフィックス表示できるようになったことから, いわゆる探索的データ解析でのグラフィックス表示の重要性を認識なさってのご発言であったかと思う．

SAS は当初から有償のソフトウェアであったが，大学・学院のご理解もいただき，たしか日本の大学で初めてセンターが導入し，学内で利用できることになった．当時マニュアルが高価でしかも英文のものしかなかったので, 槇本先生も加わって翻訳プロジェクトが動いていたことも思い出される. センターも拡大していたため, 利用者の方々個人の顔は見えにくくなっていたが, 利用者数は増えていた. 他大学の先生方から関学のデータ解析の環境がうらやましいと言われたのもこの頃である．

また，時期は少し後になると思うが，当時商学部の青木先生が中心となって，今をときめく共分散構造分析のプログラムを利用可能とされた. たしか LISREL というプログラムであったと思う．これも多分，日本の大学では先頭を切っていたのではなかろうか．

この間，筆者は SPSS，SAS についてはほとんど傍観者であり，データ解析に関わることといえば，せいぜい汎用の最小二乗法パッケージ SALS の利用環境を整備したくらいであった．個人的な研究課題として，化学物質の構造と生理活性の相関を扱い始めた頃である．これは，化学構造が説明変数で，活性値（名義属性の場合もある）が目的変数という世界である．ただし，グラフで表される化学構造をそのまま使って多変量解析というわけにはいかない．学会での主流は，化学構造から通常の変数値を生成し, 多変量解析が適用可能な世界に問題を転換することであった．筆者自身はこの傾向に逆らい，グラフ構造をそのまま扱いたいと考えていた．そのためのツールとして，人工知能研究の分野で発達した記号処理言語の Lisp を利用した．心理学の分野でも，認知科学が盛んとなって Lisp が使われ，センターのホストマシンにそのための環境を整えたことが思い出される．

このように，筆者の興味は知識処理，中でも機械学習に関係する分野にシフトしていったが, この機械学習が現在のように多変量解析と近く位置づけられるとは想像もしなかった. この時代すでに決定木が提案されており，筆者の見通しのなさが恥ずかしい．

4. データマイニング

９０年代半ばにインターネット利用の爆発とほぼ時を同じくして, ようやくデータマイニングの言葉が市民権を得た. 丁度その頃に IBM の Intelligent miner というソフトウェアが発売された. いくつかの機械学習系のアルゴリズムが組み込まれたものである. 研究助成に申請し，これも日本の大学で初めてマイニング専用のソフトウェアを導入して, センターのコンピュータラボに設置したワークステーションで動かした．この時代は研究用の計算機利用形態が，ホストコンピュータから，ワークステーション，パーソナルコンピュータへと丁度移行していた時期であり，ワークステーションの Unix 文化は，SPSS や SAS の利用者層とは重ならず，学内への波及という

点では限定されたものであった.

社会的に見ても，学会でもようやくデータマイニング関連のセッションができた時期である．80 年代には，研究の結果が出て学会で発表しようとしても，適当なセッションが無いために応用一般として扱われ，全く縁のない講演ばかりが集められたことを考えると様変わりであった．企業もデータを解析する必要性に気付き始めた頃である．しかし，先進的な企業がマイニング用のソフトウェアを導入しても，それ自体が企業秘密とされていた．

学内一般の利用者にとって，影響を与えたのは JMP の導入であろう．これもまたまた雄山先生が導入を主導され，その使いやすさも相まって現在に至るまで多くの利用者が存在する．しかし，このソフトウェアの導入は国内の大学初ではなく，慶応大学の湘南藤沢キャンパスで全員必修科目「データ分析」のツールとして導入されたことが広まったきっかけであった．データ解析の重要性が世の中に広く認知され，一握りの人間が努力して先頭を走れる時代は過ぎたということであろう．

この頃に，センターのオープン研究プロジェクト「データ解析のための知識発見システム その開発と応用」を雄山先生と立ち上げた．これは複数の企業から会費をいただき，ニュースレターを出してこの分野の紹介をするとともに，センターで開発するマイニング用のソフトウェアを利用可能とするものであった. 現在は学内でも受託研究等で企業との関わりが奨励されているが，当時産学協同は胡散臭いと見なす向きが多かった. 受託研究に比べればはるかに世間に開かれた活動であるにもかかわらず，睨まれながらプロジェクトを進めたことも今となっては懐かしい．しかし，小規模ではあってもこのようなプロジェクトが何とか成立したことは，世間でデータとその解析の重要性が認知されてきた時代と言うことであろう．

5. Predictive data miningについて

理工学部に移籍後はセンターの活動とは離れてしまった. 本稿の趣旨とはすこしずれるかもしれないが， 21 世紀に入ってよく聞くようになった２つの言葉について，最後に付言しておきたい．2000 年前後と思うが，データマイニングが一般化してきた時代に，SPSS 社（最近 IBM 社に買収されたようだが）も SAS 社もデータマイニング用のパッケージとして，Clementine や Enterprise miner の販売を始めた．その際に標語として強調されたのが predictive data mining である．「マーケティング担当者がマイニングし，消費者はこのような理由でこの商品を購入すると解釈できた, だけでは駄目である. 売り上げを上げるための施策をマイニング結果から引き出し，予測通りに売り上げが増えてこそ意味あるマイニングである」というように，筆者はこの標語を解釈している．販売担当役員ならば筆者も同じことを言うと思う．大学のまねをして担当者が論文を書いても業績向上にはつながらない，と当たり前のことを言っているだけである．しかし，筆者が引っかかるのはこの標語の裏に，「データの解釈ができなくともよい，予測が当たればよい」という姿勢が潜んでいるように見える点である．

当時はマイニングソフトウェアのパンフレット上でも, 決定木やアソシエーションルールが新しいマイニングの方法として宣伝され，解釈も予測も簡単にできるという，一見誤解を生む表現が見られた．しかし，実際のデータに適用してみると，数百の葉ノードを持つ決定木は眺めるだけでも大変であり, アソシエーションルールで出力されるルール数は下手をすると事例数よりも多くなってしまうというのが実際である．このままでは，パンフレットは嘘ばかり，そこで「ニューラルネットのように，結果を理解できなくとも予測精度が良ければそれでよし」と逃げてしまったのが predictive data mining，このように感じるのはひねくれ者の筆者だけかもしれないが，一つの見方であろう．

現在では，サポートベクトルマシンやランダムフォレストのように，より精度は高いが理解と

はかけ離れた方法が幅を利かせており, 筆者のように理解を重視する立場はますます少数派と思える. 機械学習の分野でも細かい手法の改善で精度を競う論文が多く, このままでよいのかと言われながらも, 大勢は変わらない. 自己組織化マップのようなぼんやりとした地図作りは機械学習のコミュニティの外でしか育っていない.

6. Applicability domainの問題

機械学習の論文で精度を量るためによく使われるのが, 交叉検証の方法である. 筆者の研究分野である化学構造と生理活性の相関関係を調べる場合に, この交叉検証で計算された精度は当てにならないということが常識化している. この問題を Applicability domain の問題と呼び, 10 年も前から大きな課題として浮上している.

例えば, 米国 NIH で収集した化学物質の発ガン性データを学習用データとして用い, 化合物一般に適用できる予測モデルを何らかの方法で作成したとする. 交叉検証で調べるとこのモデルの予測精度は 90%近い高い値を示す. しかし, 同じ米国の FDA が別途調査した化学物質をこのモデルで予測すると, 精度の高い方法を用いても ROC 分析で AUC が 0.5 を大幅に下回るというような悲惨な予測結果になってしまう.

この理由は, 全く出所の異なる 2 種のデータセットを学習用と検証用に用いた場合, 分子の種類は無限であるため, いずれのデータセットにも分子の選択過程で必ず何らかのバイアスがかかっているためである. 実際に化学物質の構造を詳細に調べてみると, 全く違ったデータセットを対象としており, 予測可能と考える方がおかしい場合も多い. 例えば経済時系列を予測する場合に, 石油ショック前の学習データからショック後の予測を行うと当たらないそうだが, 前記の違いは石油ショックよりもはるかに大きい. 統計的手法で補間によって予測するのは良いが, 補外は駄目という単純な話である. 化合物の活性予測では, 学習用に用いる化合物ライブラリの diversity 評価という観点から, 多くの研究がなされている. しかし, ライブラリに収容できる化合物の数に制約がある以上, 無限に存在する化合物群から十分な diversity を持つライブラリを構成することは, 実際にはなかなか難しい.

時系列的にシステムが変化する状況に対応するためには, 転移学習という呼ばれるフレームワークが提案され, 異常値の検知にもとづいて学習システムを適応させていくアプローチがよく取られている. しかし, 化学物質の活性予測には適用できない.

筆者はこの問題への対応法として,「理解できるものは予測できる, 理解できないものはやはり分からない」として, applicability domain を限定する方向で現在考えている. Predictive data mining とは異なり, 理解と予測の両立を志向する困難な途ではある. ただし, 化合物ではもちろん, プロ野球データで投球の球種や打者のヤマを予測できるかと試しているが, なかなかに難しい. この問題は考えようによっては深い問題で, そう簡単に結論が出るようなものでは無いのだろう. データ解析だけでなく, 一般の学問やビジネスで得られた知識でも同じである. 会社での経営手法が学校法人にも適用可能なのか, 西欧社会での宗教の役割が日本をはじめとする極東でどこまで敷衍できるのか, 考え出せばきりがない.

7. おわりに

筆者が関学にお世話になってからの年月だけでも, 計算機を取り巻く環境は激変した. 筆者の研究室の大学院生や卒研生の諸君に, python 言語と MySQL や R を組み合わせて, データマイニングに関する web アプリケーションを作ってもらうことが多い. いわゆるアジャイルなソフトウェア開発の方法が, 学生諸君の柔軟な知的能力の涵養にも有効と思うためである. このようなテーマは 20 年前には想像もできなかったものであり, 移り変わりの早さに我ながら驚かされる.

しかし，ASP やクラウドと言葉は変わっても，昔のホストコンピュータが別の形で姿を現しているわけで，この変わらない点も感慨深いものがある．

センターの組織が変わるのも，一抹の寂しさは覚えるが，やむをえないことであろう．しかし，データの世界が無限であり，データ解析の重要性は当分の間変わらないように思う．というよりも，かえってその重要性を増していくであろう．関学でも現在学長である杉原先生を始めとして，データ解析関連に造詣の深い先生方が多数おられる．雄山先生をはじめとする諸先生方が築きあげた，関学のすばらしいデータ解析の灯がこれからもより明るく輝き続けることを願って拙稿を閉じることとしたい．

（記憶にのみ頼って書いた原稿で，年代を含めて色々と誤っている可能性がある．また，参考文献も一切記していない．筆者に問い合わせていただけるなら，喜んで回答させていただくので，この点もあわせてご容赦いただければ幸いである．）