イベントだより

データマイニングと統計数理研究会

情報システムと人間のインタラクションや，情報システム間の通信や，取引が随所で行われるようになっている．これらのシステムは，その活動の所産である膨大なデータを，毎日・毎時・毎秒ごとに生成するようになった．同時に，それらのデータを転送し，蓄積するためのネットワークやストレージも整備された．そして，当然ながら，これらの膨大なデータを活用したいという需要が生じた．この需要に応え，大量のデータを分析する技術が1990年代に開発されはじめ，それらはデータマイニングと呼ばれるようになった．

この技術分野について議論するため，1995年には，国際会議KDD[*1]が開催された．その後も，PKDD，ICDM，PAKDD[*2]など関連国際会議も次々と誕生し，これらの会議への論文投稿数は年々増大している．国内に目を転じると，1998年頃から，現在は活動を終えたソフトウェア科学会のデータマイニング研究会が活動を始めた．同時期に，人工知能学会のKBS・FAI研究会，情報論的学習理論ワークショップ（IBIS），電子情報通信学会のPRMU研究会などでデータマイニングが取り上げられ，国内のコミュニティも徐々に拡大してきた．そこで，さらにデータマイニングについて議論するため，人工知能学会の第2種研究会として「データマイニングと統計数理研究会」を2006年に設立した．

本研究会は，人工知能分野で発展してきた機械学習や，統計学の一分野である統計的予測など，データ分析に関わるいろいろなコミュニティの交流を重視する．そのため，特に参加資格は問わず，さまざまな分野の研究者が自由に発表・参加できるようにしている．これまで，年3回のペースで開催し，そのうち1回を毎年，DMSS（International Workshop on Data-Mining and Statistical Science）として，英語によるワークショップとしてきた．これらの研究会資料は電子的に配布しており，当研究会のメーリングリストに登録することで，誰でも閲覧できる．このメーリングリストや，本研究会に関する詳細は，研究会ホームページを参照されたい．

http://sigdmsm.org/

以下，本年度に開催した，第7回研究会での討論会と，第3回DMSSワークショップについて報告する．

1. データ分析からうまれる，広がる研究と交友の輪

第7回研究会を，2008年7月23～24日に北海道小樽市にて開催した．この研究会で『データ分析からうまれる，広がる研究と交友の輪』と題したざっくばらんな討論会を催した．幹事である神嶌と樋口がそれぞれ，データ分析分野の置かれた現状についてプレゼンを行い，その後，参加者の間で意見を交換した．なお，プレゼン資料は討論会Webページ[討論会]より入手できる．

神嶌は，データマイニングに関連する研究者間の「すれ違い」について，最初に述べた．データ分析手法を研究するコミュニティは多岐にわたる．なかでも，人工知能分野の機械学習，最も長くデータを分析してきた統計，およびデータの蓄積から分析へと発展してきたデータベースの三つが，当初からデータマイニングに関係していた研究コミュニティといえるだろう．これら三つのコミュニティの関係について，文献[Zhou 03]では，図1のような洞察がなされているので，これを紹介した．研究コミュニティは，その中で価値観を先鋭化してしまうことが少なくない．この文献では，機械学習では有効性（effectiveness），すなわち，予測や分析がよりデータに当てはまることを重視していると述べている．一方，統計分野では，分析結果に理論的な保証を与え，結果ができるだけ普遍的になるようにする正当性（validity）が重んじられる．そして，データベースでは伝統的に分析のためのクエリを処理する効率性（efficiency）の向上を使命としてきた．これら三つの要素は互いにトレードオフになる傾向があり，それぞれの価値が先鋭化しすぎていて，コミュニティ間での反目も見られる．例えば，機械学習では統計分野に対し「いつも理論の都合で分布モデルを決めて，現実離れした仮定だ」などと思う．逆に，統計では「このデータについては，力ずくの調整でどうにかなっているが，全く一般性のないものだ」と機械学習手法をみなしたりする．だが，実際にデータを分析しようとすると，どの視点も無視することはできない．よって，トレードオフを考えつつ，これらの要素のバラン

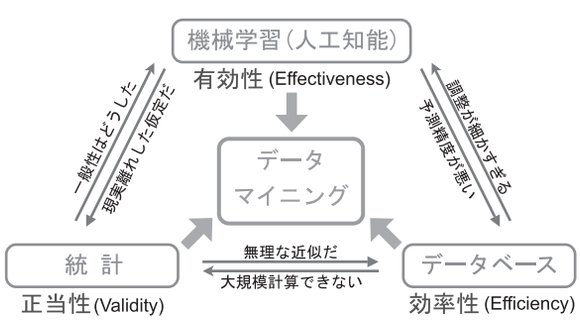

図1 データマイニングの三つの側面

---

*1 KDD: Knowledge Discovery and Data Mining

*2 PKDD: European Conference on Principles of Data Mining and Knowledge Discovery, ICDM: IEEE International Conference on Data Mining, PAKDD: Pacific-Asia Conference on Knowledge Discovery and Data Mining

スが実務的には重要になる．『データマイニング』とは，このバランスを心に留め，各コミュニティが歩み寄るためのキャッチフレーズであるともいってもよいだろう．この歩み寄りの必要性については，会場の参加者からも賛同が得られた．しかし，実際にどうすればよいかについては，特効薬はなく，コミュニティ間の人的交流の機会をつくり，時間をかけていくしかないのではないかという意見であった．

その後，データ分析手法をつくる人と，これらの手法を利用してデータを分析する人とのすれ違いについて議論した．その前に，データ分析での事前知識の役割について簡単に述べておきたい．

データを分析して結果を得るには，データだけでなく事前知識も必要になる．ここでは，データとは独立に用意した，これから行う分析についての前提のことを事前知識と呼び，統計の数理モデルや帰納的推論の背景知識といったものも含める．例として，『管』と『簡』の二つ漢字があって，次の漢字を予測するとしよう．もし全く事前知識がない，極端に客観的な状況では，ただ二つの漢字を観測した事実だけがあることになる．この状況では，たとえどんなに多くの漢字を観測しても，次は何かはわからない．今までが漢字だからといって，次が漢字であるかということさえいえない．逆に，極端な事前知識，例えば「次は『没』である」をわかっていたとしよう．このときは，今まで何を観測したかは全く無関係に，次は『没』という予測しかあり得ない．どちらも論理的には正しいが，現実の問題とはかけ離れた状況である．現実的な予測問題では，データに基づく客観性と，事前知識に基づく主観性のバランスが大切になる．例えば，『漢字の部首が重要』という事前知識を採用すれば，データから管と簡に共通する『たけかんむり』の漢字『筆』などと予測できる．この予測は，データや事前知識の一方だけから導かれたものではないので，バランスのとれたものといえよう．また，ここでは『部首』という，データの一側面を重視した．このように，予測や分類を行うときには，データのある側面を重視し，別の側面を軽視することが必然になる．このことは，醜いアヒルの子の定理（ugly duckling theorem）が意図するものである．ここで，今度は『漢字の読みが重要』という事前知識を採用したとしよう．このときは次は『完』などという予測が可能である．では『部首』と『読み』のどちらが事前知識を採用すべきであろうか？　当然，予測がより当たるような事前知識がよい．だが，現状では期待的には予測の正解率はどちらも同じになってしまう．このことは no free lunch 定理として知られている．そのため，ほかに別の根拠を見つけられないのであれば，これらの事前知識に優劣はつけられず，どちらかを直感で採用し，結果が良かったとしても，それは偶然でしかない．

前置きが長くなったが，データ分析手法をつくる人と使う人のすれ違いの話題に戻ろう．つくる人は，分析手法に思い入れがある．よって，分析手法の土台になっている事前知識が，解こうとしている問題にとって適切であるかを重視する．すると，客観性を重視した『おとなしい』結論に終わる．データ分析の研究をしていると相談をしばしば受けるが，「そんな弱い前提とこんな少ないデータで，そんな結論を出せといわれても困る」と思うことも多い．一方，使う人は，データを得るために労力やコストをかけており，目新しさのない結果では困る立場にある．いわば，使う人が立てた仮説，すなわち主観的な事前知識に基づいた結論を得たいと考える．すると，「データはあるのに，こういった結果が出ないのでは話にならない」といった不満が生じる．このように，分析手法が大切なつくる人と，データが大切な使う人はすれ違いがちである．

ではこのすれ違いを解消するにはどうすればよいだろうか？　神嶌は，つくる側からできる対策として，分析手法の今までとは違う体系でまとめられないかという案を示した．具体的には，教科書は手法の関連性からの体系付けだが，パソコンの『逆引き事典』のように，利用目的や，データの性質からの体系付けをする．これに対し，会場からは，90年代初頭には，こうした目的をもった統計エキスパートシステムについての議論が盛んだった．しかし，ドメインの事前知識をうまくモデルに組み込めないので妥当な分析ができず，統計手法の誤用を助長しただけだったという過去の事例が紹介された．バランスのとれたデータ分析をするためには，図2のKDDプロセス [Fayyad 96] を繰り返す必要があるといわれる．だが，この枠組みも，おおまかなものであり，より具体化された分析プロセスの研究も開発されるべきだろう．一方，使う人には，統計手法はブラックボックスとしては利用できないことをもっと認識してもらう必要があるのではと思う．ソートなど公理的に入出力を定義できるライブラリーとは異なり，機械学習や統計分析の目的やデータに合わせて，手法の選択や調整が必要になる．例えれば，前者が手順どおりに使えば音楽が聴けるCDラジカセであり，後者は弾き手を必要とする楽器である．よって，データを分析するときには，KDDプロセスのように反復的な作業が必要になる．このように，データの収集だけでなく，その後の分析も一貫したプロ

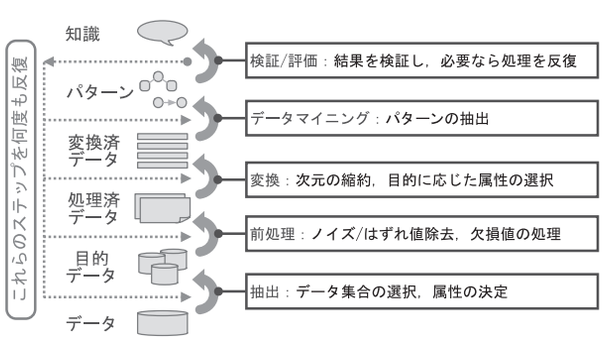

図2　KDDプロセス

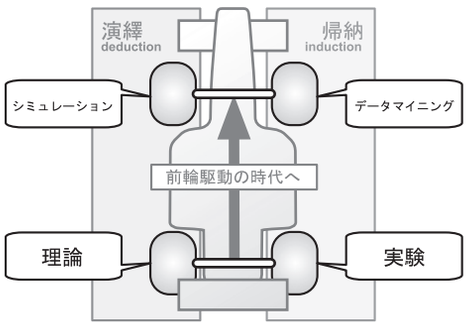

図3 科学の駆動力

セスであることを意識する必要があるだろう．そうでなければ，後から「このデータを分析すればこういう結論になるはずだが，分析するとそうならないのはなぜか？」という疑問をもつことになる．

続いて，樋口が，日本のデータ科学の置かれている苦境について述べた．科学を前進させる駆動力は，図3のように理論と実験を両輪として生み出される [佐藤 08]．理論から演繹的に導かれた結果と，実験データから帰納的に得た結果を一つのシャフトに繋ぐことで，科学の駆動力が生まれる．そして，コンピュータの登場により，理論は複雑かつ大規模なシミュレーションへと変わった．実験データも大規模になり，それを処理するデータマイニングへと移っている．こうして，前輪駆動の時代に向かっている．ここで，この駆動力が発揮されるには，この両輪が同じように回らなければならない．しかし，演繹の車輪を回すことが重視され，全体としての駆動力は十分に発揮されていない．すなわち，統計学といったデータ分析に関する学問分野は，その他の自然科学分野より重視されていない現状がある．ほかの分野での結果を導くための道具であるという側面のため，データ科学の立場は，海外でも相対的に弱いが，日本ではこの傾向はさらに強い．この背後には『物理帝国』的価値観がある．詳しくいえば，数学の公理のように，無条件に信じられるもの，ゆるがないものへの憧れである．それに対し，帰納的な統計やデータマイニングは，データを『見よう見まねする』いわば贋作の学問である．例えば，中越沖地震での柏崎原発では『想定外のGal値』という言い回しがニュースでよく聞かれた．だが，統計の観点からすれば，確率が0の事象ではない，いわば『想定していた想定外』にすぎない．ほかに，経済ニュースのコメンテータが，BSE問題で日本側が科学的な安全性検証ができないことを批判していた．しかし，科学的には危険率は0にはできないので，最後は政治的にしか結論を出せない．このようにデータ科学はゆるがないものを与えることはできない．だが，『絶対』が現実にはないこと以外『絶対』にあり得ない．これは，まぼろしに漂う演繹とは異なり，うつつに住まう帰納の宿命である．

では，この宿命を乗り越えて，帰納の車輪を回すにはどうすればよいだろうか？ それには，理論の弱い部分から実績を積み重ねることが重要である．樋口は，気象データなどの分野でデータ同化の研究に取り組んでいる．データ同化とはシミュレーションの結果を，観測データによって補正する技術である．この技術により，シミュレーションのみの場合より，実際に予測精度を上げている．このように，未知の部分があったり，不確定要素があったりするため，理論だけでは十分ではない分野が突破口である．そして，実際にデータを分析できるモデラーの育成の重要性も樋口は主張した．モデラーには，事前知識とデータとのバランスをとるため，データ分析と分析するデータの分野の両方の知識が必要となる．加えて，データ処理の実務をこなすためのプログラミング能力も身につける必要がある．こうした高度な人材を育成し，実務の場でデータ分析が活用されるようにする必要がある．

以上，今回の討論の結果をまとめてみた．今後も機会があれば同様の企画を立てたいと思う．

2. DMSS2008

第8回研究会と兼ねる形式でDMSS2008を，東工大の大岡山キャンパスにて，2008年8月25～26日の二日間にわたり開催した．DMSS（Data-Mining and Statistical Science）[DMSS] は，本研究会が開催する国際ワークショップである．今回は3回目の開催であり，DMSS2006は札幌のセンチュリーロイヤルホテルで，DMSS2007は東京の統計数理研究所で開催され，それぞれ参加者数は64人と99人であった．DMSS2008も，招待講演3件，一般発表件数16件，参加者は63名と盛況であった．国内で開催しているワークショップではあるが，国内の留学生の発表・参加により，実際に『国際』を冠するにふさわしい会議となった．以下，3件の招待講演の概要を紹介する．なお，これらの講演の講演資料は [DMSS 08] にて公開している．

最初に，津田が「New Directions in Statistical Graph Mining」の題で，大規模なグラフマイニングと機械学習手法を組み合わせた分析について述べた．生物学では，化合物やRNAの立体構造など，ベクトルの形では扱えないデータがある．そこでこれらをグラフで表現し，既存の分類や次元削減などの機械学習手法を，グラフを扱えるようにする構造化を行う．この構造化は，グラフをベクトルの形式に変換し，そのベクトルを従来通りの学習手法への入力とすることで実現する．データ中のグラフの中で頻出する部分構造をグラフマイニングの手法により列挙し，これらの部分構造パターンがグラフに存在するかどうかを特徴とみなし，これらの特徴を一つのベクトルにまとめる．ここで問題となるのは，部分構造パターンの種類があまりに多く，特徴ベクトルが大きくなりすぎて事実上計算できなくなることである．そこで，反復的で，また入力データを局所的に参照する機械学習手法がいろいろあることに注目する．こうした手法では，

機械学習とグラフマイニングによる特徴抽出とを交互に反復することで，大規模な問題を，予測精度を下げることなく問題を解けることを示した．

次の佐久間の講演「Privacy-preserving Data Mining and Machine Learning」では，プライバシー保護データマイニング（PPDM）について述べた．PPDMとは，各利用者が個人情報の入ったデータベースをそれぞれ保持している状況を扱う．このとき，各利用者は自身のデータベースの内容をほかの利用者に明かすことなく，全参加者のデータベースを集めたものを分析した結果を全参加者で共有する枠組みである．秘密を厳守する信頼できる第三者を利用する方法，任意の関数を秘密に計算する秘匿関数計算，データに乱数を加えるランダム化，そして準同型な性質をもつ暗号化を利用する暗号アプローチがある．それぞれ長所短所があるが，すべての場合に適用できるわけではないという欠点はあるものの，現実的な通信量で実行できる暗号アプローチについて詳細を述べた．さらに，k-means クラスタリングと強化学習の場合について具体例を説明した．なお，佐久間による解説が，人工知能学会誌2009年3月号に掲載されるのでぜひとも通読されたい．

最後に，Wangが「Theoretical Explanation of the Boosting Algorithms」の題で，Boostingの汎化性能に関する理論について述べた．Boostingとは，入力データからいくつもの分類器を生成し，それらの予測結果を統合して最終結果を得るアンサンブル学習の一つである．直感的には，現状で最も分類しにくい，苦手なデータに注目した分類器を，逐次的に加えることで予測精度を向上させる．このBoostingでは，分類器をある程度加えると経験誤差が0になる．だが，その後も分類器を追加し続けると経験誤差は0で変わらないが，汎化誤差も小さくなる現象がしばしば見られる．この現象は従来はマージンが大きくなるためであると説明されてきた．しかし，最小マージンをより大きくできるように設計したBoostingであるarc-gvを適用すると，汎化誤差が悪くなるという，従来の理論に反する実験結果が得られた．そこで，新たに平衡マージンと呼ぶ概念を示し，最小マージンではなく，平衡マージンを大きくすることで汎化誤差を減らせることを示した．

以上が，第7回とDMSS2008の活動報告である．次の第9回は2009年3月に京都にて開催する．また，研究会に勉強会の要素を加えるreview発表も新たに始めるので，ぜひとも発表・参加されたい．また，本研究会の資料を購読できるメーリングリストへの参加も歓迎する．

◇ 参 考 文 献 ◇

[DMSS] 過去に開催したDMSSのホームページ一覧，http://sigdmsm.org/dmss.html

[DMSS08] DMSS2008招待講演一覧，http://sigdmsm.org/dmss2008/italk.html

[Fayyad 96] Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P.: From data mining to knowledge discovery: An Overview, in Fayyad, U. M., Diatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. eds., *Advances in Knowledge Discovery and Data Mining, chapter 1*, pp. 1.34, AAAI Press/The MIT Press（1996）

[佐藤 08] 佐藤忠彦，樋口知之：動的個人モデルによる消費者来店行動の解析（討論付），日本統計学会誌，Vol. 38，No. 1，pp. 1-38（2008）

[討論会] 討論会『データ分析からうまれる，広がる研究と交友の輪』，http://sigdmsm.org/007/discussion.html

[Zhou 03] Zhou, Z.-H.: Book Review: Three perspectives of data mining, *Artificial Intelligence*, Vol. 143, pp. 139-146（2003）

〔神嶌　敏弘（産業技術総合研究所）〕