特　集「研究会総覧」

データマイニングと統計数理研究会 (SIG-DMSM)

Special Interest Group on Data Mining and Statistical Mathematics

神嶌　敏弘　産業技術総合研究所，データマイニングと統計数理研究会主査
Toshihiro Kamishima　National Institute of Advanced Industrial Science and Technology (AIST)
mail@kamishima.net, http://www.kamishima.net/jp/

データマイニングと統計数理研究会（SIG-DMSM）はデータマイニングを実現するための機械学習，統計数理，およびその他の各関連分野の理論・手法・技術に関する基礎から応用までの幅広い研究課題を対象とした研究会である．

統計は，データを扱う手段としては最も長い歴史をもつ．回帰分析やFisher判別分析などの基礎的な統計的予測手法は，現在でも重要である．加えて，一般化線形モデルや粒子フィルタなどの新たな手法も活用されている．

機械学習は，人間が日々の実体験から得られる情報の中から，後に再利用できそうな知識を獲得していく過程を，コンピュータにおいて実現したいという動機から生じた．そして，数値・文字・画像・音声など多種多様なデータの中から，規則性・パターン・知識を発見する技術として利用されている．

当初は命題論理などに基づく記号的知識の学習が主であった機械学習も，徐々に確率的な枠組みを吸収してきた．一方，線形モデルが中心であった統計的予測においても，カーネルを使ったサポートベクタマシンなど機械学習と共通したモデルが採用されるようになった．このように両者は，現在では密接に関係している．同時に，図に示すように，データの規則性を扱う情報理論や計算論的学習理論，実際のデータ処理計算に必要となるアルゴリズム論や統計物理，大規模データを扱うためのデータベースや並列計算技術などとも密接な関連がある．

一方，実世界においては，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・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・毎秒ごとに生成しつつある．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てきた．この状況下で，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，機械学習や統計的予測の技術を用いて，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．

図1の上部に示すように，機械学習・統計・データマイニングなどは，データを扱うあらゆる分野で利用さ

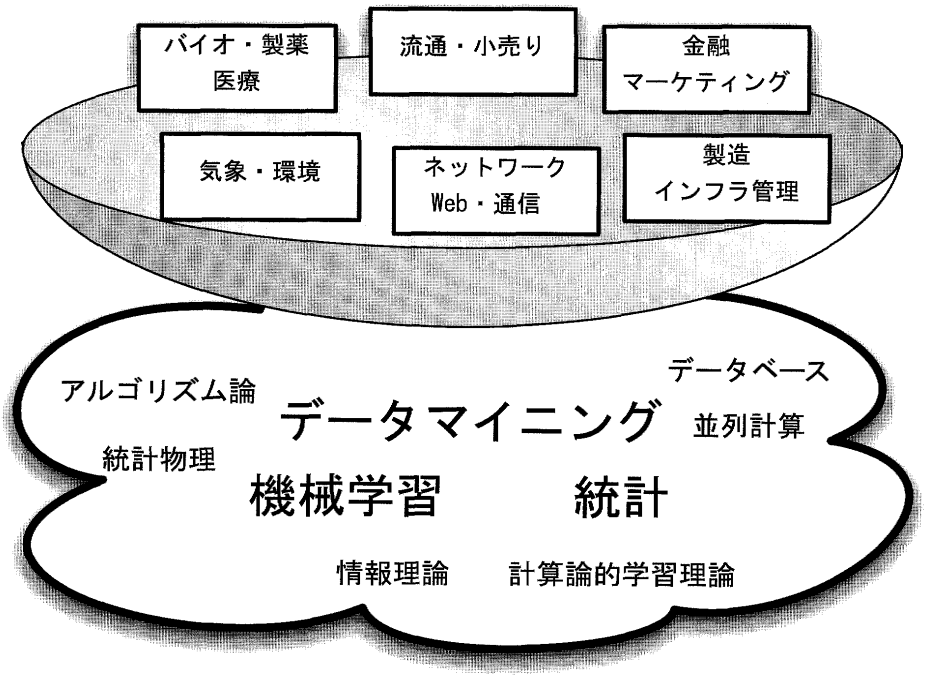

図1　データ科学とその関連分野

れている．バイオ・製薬・医療への応用としては，遺伝子そのものやその活動を見るマイクロアレーデータの分析，データに基づく医療（EBA: Evidence Based Medicine）がある．マーケティング・金融への応用においては，小売りデータ解析に基づく顧客購買傾向の予測や，融資記録データに基づく与信評価などがあげられる．気象・環境の変化予測，災害対策，製造ラインや電力網などの効率的な運用にもデータマイニング技術が使われている．機械学習は，通信・ネットワークでも効率的な運用を実現するために利用されるとともに，これらの上で動作するWebでは，製品やサービスの効率的な広告や推薦を行うために欠かせない技術となっている．

こうした状況を背景に，機械学習と統計的予測の双方の技術を基盤とするデータマイニングについて，議論や情報交換を行うため，人工知能学会の第二種研究会として，本研究会を2006年度に設立した．では，これまでの12回の研究会を順に振り返ろう．

第1回：2006.7.11，統計数理研究所

キックオフということで，招待講演4件という豪華な回であった．有村は不均一で弱く構造化された半構造データからのパターン発見，狩野は観測結果から因果関係を推定する因果推論，福水は非線形問題を線形問題のように扱えるカーネルの理論，そして上田は自然言語処理などで応用が広がっているノンパラメトリックベイズについて講演した．

第2回：2006.9.25-26，札幌

科研の半構造データマイニングが主催するDMSS（Data-Mining and Statistical Science）ワークショップに協賛する形で開催した．これは開催は国内だが，発表の多くは英語で行われた．

第3回：2007.2.27-28，兵庫県立大学

佐藤はクラスタリングについて，鈴木は分析結果の興味深さが分野によることを講演した．

第4回：2007.7.25-26，旭川

水田による次元削減についての講演が行われた．

第5回：2007.10.5-6，統計数理研究所

この回より本研究会の主催として2回目のDMSSワークショップを開催した．行列分解を使った協調フィルタリングについてのSmola，バイオインフォマティクスに関するSheridan，およびグラフカーネルについてのVertの講演があった．

第6回：2008.2.28-29，大阪大学

病院情報システムについて津本が講演した．

第7回：2008.7.23-24，小樽

討論会『データ分析からうまれる，広がる研究と交友の輪』を行い，分析手法をつくる側と使う側のすれちがいと，演繹的な理論が本物とされ帰納的なデータ科学が軽視される現状について述べた．

第8回：2008.9.25-26，東京工業大学

3回目のDMSSであり，プライバシー保護データマイニングの佐久間，ブースティングのWang，大規模グラフカーネルの津田の講演があった．

第9回：2009.3.3-4，京都

劣モジュラ最適化について岩田が講演した．

第10回：2009.7.7-8，京大会館

4回目のDMSSで，データ構造ZDDを用いたマイニングの湊とマーケティングでのマイニングについての矢田の講演があった．

第11回：2009.10.18，九州大学

機械学習・学習理論で10年以上の歴史をもつIBISワークショップと連続開催した．サーベイ，自身のまとめ，ポジションペーパなどのreview発表の特集であった．

第12回：2010.3.29-30，統計数理研究所

日本学術会議の分科会との共同企画として，北川はデータが活用される知識中心の社会のビジョンについて，中島は「スマートシティはこだて」プロジェクトについて講演した．

以上のような活動を続けてきた本研究会だが，第12回を最後に，信学会のIBIS研究会と合併し，次年度より信学会にて『情報論的学習理論と機械学習研究会（IBISML）』となる．この度，AI学会を離れることとなったが，機械学習の応用分野の研究会との共催・連続開催などを通じAI学会と連携したいと考えている．その際には，ぜひとも参加されたい．