特集

最新！データマイニング手法

New Frontier of Data Mining Methods

0 編集にあたって

鈴木 英之進（横浜国立大学大学院工学研究院）
suzuki@ynu.ac.jp

鹿島 久嗣（日本 IBM（株）東京基礎研究所）
HKASHIMA@jp.ibm.com

1989年，人工知能の国際会議IJCAIの併設ワークショップとして記念すべき第1回のKDD Workshopが開催された．その名の由来である「データベースからの知識発見(Knowledge Discovery in Database)」というコンセプトは，従来，大量のデータを相手に仮説検証型の演繹的推論を効率的に行う技術を研究していたデータベースの研究者たちには，計算機が仮説を自動的に生成するという帰納的推論という視点を与えた．一方，人工知能の一分野でデータからの帰納的推論を研究していた機械学習の研究者たちには，メモリに入りきらないほどの大規模なデータを効率的に扱わなければならないというアルゴリズムのスケーラビリティという視点を与えた．この結果，2つの研究分野が結びつき，そこに統計学者たちも加わることによって，データマイニングの原型が形成された．

そして，1990年代初め，IBMアルマデン研究所のAgrawalらによってデータベースからの相関ルール発見に関する一連の論文が発表された．これらの核であるAprioriアルゴリズム自体は，データの中に頻出する値の組合せパターンをすべて数え上げるというシンプルなものであったが，構造化された解空間においてこれを効率的に行う手法と，数え上げたパターンを組み合わせることで「AならばBである」というかたちの相関ルールと呼ばれる局所的に成立するルールを導くというアイディアは，データマイニングに相関ルールという新たな要素技術をもたらし，多くの研究者のみならず，実務家をもこの分野に惹きつけた．

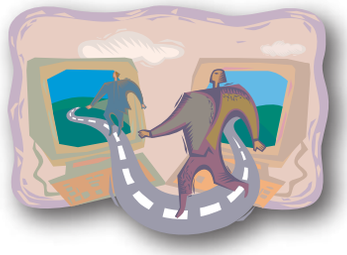
以降，参加者の増加とともにKDD Workshopはその規模を拡大してKDD Conferenceへと成長し，また，関連する国際会議やその参加者も順調に数を増やし，データマイニングは現在に至るまで活発に研究がなされる一大分野を形成している．

さて，本特集は，現在データマイニングの分野で活躍している研究者らによる最新動向の解説である．執筆者らは人工知能，データベース，統計学の三分野にまたがり，ベテランから若手までバラエティに富んでいる．取り上げられるトピックは，大きく「量」「形」「応用」の3つの視点から分類することができるだろう．

まず1つ目のカテゴリーはデータの「量」に注目した解説である．

有村らによる「データストリームのためのマイニング技術」は，データストリームと呼ばれる，取引記録や通信記録，センサから収集されるデータなど，時間的に変化する大量の電子化されたデータの流れをほぼリアルタイムで解析する手法についての解説である．時々刻々と大量に押し寄せるデータの波をいったんデータベースに溜め込んでから解析するというのはほぼ不可能であるが，データ系列を逐次的に処理することで，効率的にデータストリームを解析するための一連の手法を紹介している．

鈴木による「データスカッシング－逆転の発想によるスケールダウン戦略－」は，データベースに蓄えられた大量のデータを，データの重要な特徴を失うことなく，上手に「変換する」ことによって，マイニングアルゴリ



ズムを高速に走らせるという，データスカッシングという手法についての解説である．通常の機械学習アルゴリズムはメモリに載りきらないほどのデータをそのまま扱うようには設計されていない．データスカッシングは，学習アルゴリズムを変更することなく，逆にデータをうまく要約することによって大量のデータを扱えるようにするという逆転の発想に基づく．

2つ目のカテゴリーは，データの「形」に注目した解説である．

従来，解析の対象となるデータは，関係データベースに納まるような項目と値のペア，言い換えればベクトルの形で表現できるということを暗黙の了解としてきた．しかしながら近年，ベクトル形式ではうまく表現できない複雑な構造を持ったデータに対しても，同様の解析を行おうという試みが盛んに行われている．それらの例として，テキストデータやDNAなどの配列によって表現したほうが都合の良いデータ，HTMLやXMLなどのように木によって表現できるデータ，あるいは化合物やWebのリンク構造などのようにグラフの形で表現できるデータなどがある．

鷲尾による「グラフベースデータマイニングの基礎と現状」は主に，彼らが世界で初めてグラフデータに対する拡張を行ったパターン数え上げ手法の基礎と，その発展の最新動向を紹介している．彼らが立ち上げた研究分野は理論と応用の両面から世界的に注目されており，主催する国際ワークショップも非常に盛況と聞く．

鹿島による「カーネル法による構造データマイニング」は，近年注目されているサポートベクトルマシンなどのカーネル法と呼ばれる手法によって，構造を持ったデータを解析するためのアプローチを紹介している．カーネル関数と呼ばれるデータ間の類似度をうまく設計することで，構造の持つ特徴を捉えながら，効率的に計算を行う手法を解説している．

3つ目のカテゴリーは新しい「応用」に注目した解説である．

近頃，世間ではコンピュータウィルスや情報漏えいなどの，情報セキュリティにまつわる問題が盛んに取り沙汰されている．山西らの「統計的異常検出3手法」では，これらのセキュリティの問題や，あるいはシステムの異常や障害の自動検知の問題において不可欠な技術である異常検出に焦点を当てている．彼らの統計的手法に基づく外れ値検出と変化点検出技術は，リアルタイムでモデル構築と異常検出を行うことができ，不正侵入，ウィルス，なりすましなどの対策として有効である．

近年，Webの掲示板や，オンラインでのアンケート，電子メールなど，電子的な形でテキストデータが簡単に大量に入手できるようになっている．工藤らの「自然言語処理におけるマイニング技術の応用」は自然言語処理のタスクに最新のデータマイニング技術を応用することで，テキストデータを活用する試みを紹介している．機械翻訳における対訳表現の自動抽出や，Web上での評判分析などにおける文書分類などのタスクをデータマイニング手法に基づいて解釈しなおし，改良を加えることで適用に成功している好例である．

豊田らの「大規模Webアーカイブからのデータマイニング」は，おそらく世の中で最も大きなデータベースの1つであろうWebのダイナミクスに注目した解説である．彼らは，Internet Archiveに代表される非常に大規模なWebのアーカイブを解析することで，Webにおけるトピックの移り変わりや発展過程を捉え，視覚化するという非常に野心的な試みを紹介している．

データマイニングの出現から十余年がたち，その概念は広く世の中に浸透したといえるが，データの洪水の中から，データに潜むエッセンスを捉え，活用するというデータマイニング技術は，今後の情報化社会においても依然として，いや一層重要な技術としてあり続けるであろう．本特集によって読者がデータマイニング研究の最前線における新しい潮流の一端を感じ取っていただければ幸いである．

（平成16年12月6日）