3 日本・アジアにおけるデータマイニングコンテスト

鈴木英之進（横浜国立大学工学部電子情報工学科）
suzuki@dnj.ynu.ac.jp
津本　周作（島根医科大学医療情報学講座）
tsumoto@computer.org

日本発　専門家指向型コンテスト

データマイニングコンテストは，複数の参加者が共通のデータ集合を各々の手法で解析し，それぞれのデータ集合に与えられた課題を解決し，得られた結果を比較評価する試みである．カーネギーメロン大学のMitchellは，データマイニングにおける研究課題の1つとして決定（decision）の最適化を挙げ，これは予測（prediction）の最適化に比較して困難であると述べている[5]．データマイニングの問題解決過程は，FayyadらのKDD（Knowledge Discovery in Databases）プロセスモデル[3]によれば選択，前処理，変換，抽出，および解釈の5段階に分けられ，複数手法の反復適用を必要とする．データマイニングコンテストは，予測問題に比較して複雑なデータマイニング問題における種々の手法を比較評価する上で有意義な試みであると考えられる．

アジアにおけるデータマイニングコンテストは日本から始まった．日本では「共通データからの知識発見」[7]の名称のもと，人工知能学会において研究会や全国大会の一環としてまず4回のコンテストが行われた．これらのコンテストは，著者の1人である津本がとかく基礎研究に偏りがちである人工知能研究の現状を憂慮し，次の目的を念頭に開催した．

- 人工知能側からの独自手法の開発：データマイニングは大量情報からの有用知識発見を目的とするため，データベース，統計学，および人工知能研究の境界領域研究と位置付けられる．現在人工知能側から発表されたデータマイニング手法は，従来の機械学習の枠を超えないか，そこに統計学的手法を採り入れたものがほとんどである．データマイニングコンテストの開催により，これら3分野における人工知能の独自性をうまく活用した手法の提案を促したい．
- 実問題指向の手法開発：たとえばFisherがIrisデータを分類するために線形判別関数[4]を導入したように，統計学は実世界のデータから出発して新たな手法を開発してきた．いわゆる“toy problem”ではない実問題をデータマイニングコンテストで用いることにより，問題点の議論と分析を通してこれらを解決する実問題を指向した手法の開発を促したい．
- 評価方法の提案：予測手法が正答率や最小二乗誤差などの客観的基準によって評価されるのに対し，データマイニング手法は発見された知識の有用性によって評価される．知識の有用性は問題によっては予想利益などの客観的基準で測られるが，通常は知識を利用する人間の主観的基準で測られる．データマイニングコンテストでの結果評価を通して，長年「専門家の知識」の実装に取り組んできた人工知能研究の独自性を活かした評価方法を提案したい．

津本は医用統計学・人工知能研究者であるとともに医師でもあることから，データマイニングコンテストに医療データを提供し発見された知識を評価することができる．医療分野は商業分野などと異なり利益などの唯一絶対の客観的な評価基準が存在しない．たとえばある検査値と症状の関係は医療行為に直接影響は及ぼさないが，症例の深い理解につながる可能性があるために仮説としてきわめて興味深い場合がある．本稿で解説する日本・アジアにおけるデータマイニングコンテストは，米国で開催されたKDDカップ1997，1998，1999に比較して仮説生成を重視する専門家指向型コンテストとしての色彩が濃い．

なお「共通データからの知識発見」の中心メンバたちは，2000年4月に国際会議の付設ワークショップとして国際コンテストKDD Challenge 2000[6]を開催した．KDD Challenge 2000は，先行して行われた4回のコンテストで用いられた発見問題に加え，変異原性データも用いた意欲的な試みである．KDD Challenge 2000の後，やはり人工知能学会での研究会の一環として，アミノ酸

配列データを用いた第5回目の「共通データ集合からの知識発見」が行われた[9].

共通データからの知識発見

人工知能学会において研究会や全国大会の一環として現在までに5回のコンテスト「共通データからの知識発見」[7), 9)]が行われた．これらのコンテストを，用いたデータや発表件数などとともに**表-1**に示す．

表-1よりコンテストの参加者を招待者に限定した2回を除けば，髄膜炎データとアミノ酸配列データを用いた回は参加者が多く，膠原病データを用いた回は少ないことが分かる．これは主にデータの難易度に原因があると考えられる．コンテストで用いたデータの構造を**表-2**に示し，与えた初期課題とともに簡単に説明する．

- 髄膜炎データ：首都圏の某都市の3次救急指定病院において昭和54年から平成4年までに入院治療した140人の髄膜脳炎（meningoencephalitis）患者に関するデータ．内容は入院時の病歴，検査所見，専門医の鑑別診断，治療内容，治療後の経過，および転帰．欠落値が1つの属性だけに現れることからも分かるように，前処理をすべて終えたデータに分類される．初期課題は1）診断に重要な因子，2）原因菌探索に重要な因子，および3）後遺症の有無に関して重要な因子の特定である．
- 細菌検査データ：某病院において1994年に行われた細菌検査約20,900件に関するデータ．内容は細菌培養検査の結果，検体の情報，血液検査，および尿検査など．検体の病名が3個の属性に順序を考えずに記載されていることからも分かるように，前処理を行っていないデータに分類される．初期課題は1）原因菌の（+）と（−）に関連する属性，および2）各種抗生物質の抵抗性と感受性に関連する属性の特定である．
- 膠原病データ：某大学医学部附属病院第二内科膠原病外来に数年以上通院し，診断・治療・経過の観察が行われている約1,000人の患者に関するデータ．膠原病（collagen disease）は自らの免疫機構が組織を攻撃するために発現する疾患であり，その合併症には小血管が閉塞してしまう血栓症（thrombosis）などがある．データの内容は検体の情報，各種検査値，および各種抗体の情報．前処理を終えたデータと前処理を行っていない時系列データから構成される．初期課題は1）血栓症の診断に有効な属性と時間的パターン，2）各膠原病の正確な診断を可能とするパターン，および3）各膠原病に特徴的なパターンの特定である．

時期	会議名（人工知能学会）	データ集合	発表件数	備考
1998年6月	全国大会	髄膜炎	5	招待だけ
1999年1月	KBS研究会	髄膜炎	11	
1999年6月	全国大会	細菌検査	4	招待だけ
1999年9月	FAI／KBS合同研究会	膠原病	4	
2000年9月	KBS研究会	アミノ酸配列	9	

表-1　共通データからの知識発見における5回のコンテスト

名称	ファイルのレコード	属性数（ID除く）	備考
髄膜炎	患者140人	38	
細菌検査	細菌検査20,919件	162	ほぼすべて欠落値あり．一部時系列．値の種類が膨大な属性と集合属性
膠原病	患者1,240人	6	
	特殊検査を受けた患者806人	12	
	検査57,542件	42	ほぼすべて欠落値あり．時系列
アミノ酸配列	アミノ酸配列34本	237	

表-2　共通データの構造

- アミノ酸配列データ：東北大学生物工学専攻において実験で計測された，ニワトリリゾチームを抗原とする抗体に関する34本のアミノ酸配列に関するデータ．抗原抗体反応は，免疫反応において最も重要な反応の1つであり，抗体が抗原に結合することで抗原の持つ機能などを失わせる過程を表す．データの内容はアミノ酸配列，結合定数，および熱力学変化などである．背景知識を導入していないデータに分類される．初期課題は，配列と結合定数あるいは配列と熱力学変化に関する，相関関係か回帰式の特定である．

コンテストにおいてはデータマイニングのインタラクティブ解析としての側面を重視し問題設定と解析方針は自由で発見結果を専門家が評価する方式を採用した．本コンテストが医療分野を重視し医療においては専門家に対する仮説生成が重要であることより，この設定は妥当だと考えている．ただし解析の手がかりを与えるため，上に示す初期課題も与えた．

KDD Challenge 2000

KDD Challenge 2000[6)]は，第4回太平洋アジア地域知識発見とデータマイニング国際会議（PAKDD-2000）に併設されて2000年4月18日に行われたワークショップである．出題問題は前章で紹介した髄膜炎データ，細菌検査データ，および膠原病データと，次に説明する変異原性データである．

- 変異原性データ：1991年当時カリフォルニア州ポモ

コンテスト時期	データ	発見知識形式と参加者数	用いられた発見手法
1998年6月	髄膜炎	ルール3，決定木2，属性1	離散化＋相関ルール，対話型決定木，KDDプロセス，ラフ集合論に基づくルール，統計的検定と多変量解析
1999年1月	髄膜炎	ルール8，決定木3，属性1，事例1，クラスタ1	対話型決定木，GP＋決定木，多戦略のメタ学習，ラフ集合論に基づくルール，トップダウンビーリング，離散化＋相関ルール，領域分割＋2次元数値属性相関ルール，SVMによる事例，例外ルール，カスケードモデル，可視化手法＋階層型クラスタリング
1999年6月	細菌検査	ルール3，決定木1	例外ルール，対話型決定木，多戦略のメタ学習，離散化＋相関ルール
1999年9月	膠原病	ルール4	GA＋帰納的論理プログラミング，多戦略のメタ学習，可視化＋決定木＋ルール，グラフ構造型化＋相関ルール
2000年4月	髄膜炎	ルール2，属性のグラフ1	決定木＋属性可視化＋ルール，多戦略のメタ学習
	細菌検査	ルール1	例外ルール
	変異原性	ルール3	グラフ構造データからの相関ルール，カスケードモデル，統計的検定に基づくルール
2000年9月	アミノ酸	ルール6，属性1，回帰モデル2，属性間の相関1，クラスタ1	相関ルール＋GP，ラフ集合論に基づくクラスタリング＋属性特定，カスケードモデル，NNからの法則発見，離散化＋ルール，特異ルール，多戦略のメタ学習，線形回帰＋相関解析，一般化線形モデル

表-3　参加者が用いた手法の分類．ただし，GP，SVM，GA，およびNNはそれぞれ遺伝的プログラミング，サポートベクタマシーン，遺伝的アルゴリズム，およびニューラルネットワークを表す．なお複数種類の知識を発見する参加者がいるため，発見知識形式の合計数は表-1の発表件数に一致しない場合がある．

ナ大学のDebnathらが発表した230個のニトロ化合物の変異原性（mutagenicity）に関するデータ．変異原とはDNA配列に突然変異を起こす性質を持つ化学物質等を指し，発ガン活性と高い相関を持つことが知られている．データの内容は化合物の構造式，疎水性，最低空軌道のエネルギーレベル，および生理活性値である．構造を含むデータに分類される．課題は1）活性を発現させる部分構造の発見，2）活性を抑制する部分構造の発見，および3）活性と他の特徴量との間の定量的相関関係の導出である．このデータは表形式部分とグラフ構造を指定する2個のファイルから構成され，230個の化合物に対するデータが記されている．なお，構造式から商用ソフトウェアにより抽出した表形式の構造特徴ファイルも別に与えられており，上記目的に利用することができる．

KDD Challenge 2000での発表件数は6件だった．なおコンテストで使用されたデータは，文献6）からダウンロードできる．

コンテストはるつぼ

参加手法

データマイニングコンテストでは上述のようにさまざまな種類のデータが用いられるが，参加者の手法も多種類となっている．用いられた手法を分類したものを表-3に示す．表より発見知識の形式としてはルールが過半数を占め，他を引き離して多いことが分かる．これはルールが可読性に優れることもあるが，データの一部だけを説明するためでもあると考えられる．次節で説明するが専門家の興味を引く発見知識は，領域知識に照らし合わせてある程度妥当であるものの，意外性を含む場合が多い．ルール発見では説明するレコード集合が一意に定まらないため種々のルールが発見され，その中では妥当性と意外性の両方を有するものも存在する．このためルール発見は専門家指向のデータマイニングに適していると考えられる．一方決定木は，機械学習において最もよく研究されている分類モデルの1つであるためか，最初2回のコンテストにおいては計5回用いられてルールの次に多い．しかし後のコンテストではほとんど用いられておらず，これは分類モデルがレコードすべてを説明するためであると考えられる．決定木はルールに比較して妥当性と意外性の両方を有することが難しく，コンテストを重ねるうちに次第に用いられなくなっていった可能性がある．

表よりコンテストでは機械学習と統計学における種々の手法が用いられたことが分かる．これらの中には時系列データをグラフ構造に変換しそこから相関ルールを発見する方法など，課題問題に応じて独自に開発した手法も少なからずあった．以上から人工知能独自の手法の開発を促す目的は達成されつつあると考えられる．参加者の約半数が複数個の手法を組み合わせたことは，データマイニングが複数手法の反復適用を必要とするというFayyadらのKDDプロセスモデルの実証となっている．なお同じ手法が複数のコンテストに現れるのは，該当する参加者が複数回参加したためである．ただしその場合でも，参加者は同じ手法を与えられたデータに応じて改良することが多い．このことより実問題指向の手法を開発してもらう目的も達成されつつあると考えられる．

発見知識

各回のデータマイニングコンテストでは重要な発見や知見が得られた．ここでは専門家の興味を特に引いた発見を教訓として整理する．なお紹介する手法はそれぞれ興味深い発見を成し遂げており，ここで紹介する内容だけに基づいてその価値を判断しないように注意されたい．

①興味深い知識は問題領域において少し意外である．

コンテストの開催前より，一見奇異に思われ過適合と思われるルールも再検討することにより，興味深い発見につながることがあると分かっていた．コンテストにおいては，このことの正しさを示す例が多数得られた．新美・田崎らによる遺伝的プログラミングを用いたルール発見では，必ずしもデータをカバーする率は高くないが仮説となり得るルールが得られた．このことは標準的な決定木学習手法C4.5が，多くの例をカバーし専門家にとってはきわめて妥当だが興味深くないルールを出力したことを考えると示唆深い．鷲尾らの離散化では，クラス情報を用いるものは医学的に妥当だが，クラス情報を用いないものは保守的すぎて医学的には使いものにはならなかった．専門家は典型的な例に関する結果よりも「境界例」に関する情報が欲しいと思っており，妥当すぎる知識を発見することはかえって有効ではない場合が多い．山口らのルール発見でも，妥当性と興味深さのトレードオフが観察された．

②領域知識の使用には長所・短所がある．

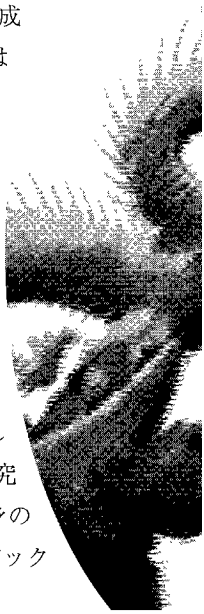
知識発見においては，領域知識を用いたために興味深くない知識を除外できる場合と，興味深い知識をかえって発見できない場合がある．鈴木らが発見した例外ルールの1つは，抗生物質の有効性が病院の部署によって大きく異なることを示し興味深いと思われたが，（結果）→（要因）を表していた．このような例は領域知識を与えることにより防げたと考えられる．一方コンテストが開催される前，津本らはラフ集合論に基づくルール発見に基づき，慢性疾患と強い関連性を持つ要因を特定した．この要因は，専門家が領域知識に基づき解析属性を限定していたために未発見だった．なお欧州のデータマイニングコンテストでは，Spenkeらが独自に開発した可視化ツールに基づき領域知識の欠如を補い専門家に酷似する手順で前処理を行い，専門誌に掲載されるほどの知識を発見した[1]．この例は前処理において基本的な手順を踏めば，あたかも領域知識があるがごとく解析が可能であり，そのような前処理をうまく行えば専門医にも予想できない知識が抽出され得ることを示している．

③構造化されたルールの発見は有効．

鈴木らの例外ルールや岡田らのルールは，データにおいて一般的で正確なルールに付加情報が加わると結論が変わる構造を持ち，発見知識としてきわめて魅力的である．鍾らのルールも条件に応じて結論部が変わる構造を持っており，発見知識として有用である．これらの例は，独立したルールよりも互いに関連し合う構造化されたルールの方が明確な意味を持つため発見対象として興味深いことを示している．

④データマイニングコンテストは研究課題の宝庫．

たとえば膠原病データは，計測間隔がきわめて不規則な高次元時系列データ集合であり，ほぼすべての属性に欠落値がある．さらにこの領域では長期間にわたるパターンが重視され，重要な解析目標である血栓症が少数クラスとなっている．鷲尾らの時間データをグラフ構造に変換してそこから相関ルールを発見する手法も，山口らのメタ学習に基づき複数の発見手法を組み合わせて解析プロセスを自動合成する手法もきわめて興味深いが，現時点ではこれらの問題点を完全に克服できたとはいえない．コンテストのデータ集合はデータマイニングや知識発見の研究分野に多数のチャレンジを課していると見なせる．



参加者

データマイニングコンテストは，さまざまなバックグラウンドと興味を持つ参加者が集まる「るつぼ」でもある．本稿で紹介したコンテストへの参加者は主に機械学習研究の出身であるが，統計学やデータベース出身の者も存在し，計算機科学とは異なる領域のバックグラウンドを持つ者も多かった．

KDD Challenge 2000の開催組織は，データマイニングの研究者から構成されるプログラム委員会と出題領域における専門家から構成される評価委員会の両方から成り立っていた．プログラム委員の関心は，たとえば変異原性の説明など出題された領域固有の問題を解くことと，たとえばグラフ構造からの相関ルール発見などデータマイニングにおける問題を解くことの2つに分かれていたと思われる．ただし故Zytkow教授は科学史研究のバックグラウンドを持ち機械発見に大きく貢献してきた経歴から，データマイニングコンテスト参加者同士のシナジーにより知識が発見されていく過程に関心を持っていた[8]．評価委員のコンテストに対する態度は，どちらかといえば懐疑的なものもあったが，仮説生成としては有用と見るものや積極的に期待するものもあった．もっとも大多数の専門家は，自らの研究に何らかの

有意義なフィードバックを得ることができたと認めている．

発見の評価法？

科学史を振り返るまでもなく，発見の評価は困難である．仮説生成を目的とするデータマイニングコンテストにおいてはまず，発見知識の評価指標を設定する必要がある．データベースからの知識発見は，妥当で新規性があり，有用で理解可能な知識を求めることを目的とする[3]．これらの評価指標，すなわち妥当性（validness），新規性（novelty），有用性（usefulness），および理解可能性（comprehensibility）は正確には定義できず，専門家でも明文化できないものである．したがって知識発見では発見知識を，これらの指標の代わりに正答率で評価する場合も多い．もっとも実世界の学習において正答率は評価指標の1つに過ぎず，専門家に軽視される場合も多いことが報告されている[2]．実際には妥当性，新規性，有用性，および理解可能性などいくつかの評価指標を決め，専門家に判断してもらうのがよいと考えられる．ただしこの評価法は専門家に比較的大きな負担をかけることが分かってきており，決定的な方法というわけではない．

発見知識の評価法が定まったとしても，知識発見手法の評価方法が決まったわけではない．特にデータマイニングコンテストには，知識発見手法の理解を目的とする手法指向の立場と，未知で有用な知識の発見を目的とする問題指向の立場がある．これらの立場は相反する場合もあるので，コンテストの目的を明確にすべきである．たとえば，100個の発見ルールのうち99個が比較的有用である手法と，100個の発見ルール中99個は間違っているが1個がきわめて有用である手法では，どちらが優れているだろうか．手法指向の立場では前者が優れており，問題指向の立場では後者であるといえる．

最後にデータマイニングプロセスが複数の段階から構成され，知識発見手法は通常それらの一部分だけを対象とすることにも注意しなければならない．投稿手法を知識発見の成功に関連付けることにより知識発見手法を標準化するためには，明確な規則の元での網羅的な実験が必要である．当時ダイムラー・クライスラーのWirthは PKDD'99 での招待講演で，KDDプロセスモデルの標準化が持つ重要性について強調した．知識発見コンテストは将来，投稿手法をプロセスモデルの各段階に分類し，可能な組合せを網羅的に試みて発見知識を系統的に評価する形態に移行する可能性もある．

発見手法や発見結果の評価は重要な問題であり，有用性が明確な形で記述できないので適切な設定が必要である．発見手法や発見結果の評価法の確立は，冒頭で述べた他の2つの目的に比較していくつかの知見が得られた段階であり，まだ達成されたとはいえない．

今後の展望

日本・アジアにおけるこれまでのデータマイニングコンテストは，仮説生成を重視する専門家指向のコンテストとして，結果の評価において主観的評価を重視してきた．このことは人工知能側からの独自手法を活かした手法の開発や評価方法の提案に確かに貢献したと考えられる．ただし冒頭でも述べたが，データマイニングにおいては予測ではなく決定の最適化が重要である．今後は，たとえばファンド投資やプラント運転など，発見知識の価値が短期間で明確に分かる領域でのコンテストも必要である．別の領域としては，参加者の増加や社会への影響を目的としてアミューズメント分野や通信分野などが考えられる．我々は今後のコンテストにおいても手法と領域の両方で着実に成果を挙げ，応用領域を広げていかなければならない．

謝辞　これらのコンテストは，多くの人の貢献がなければ成功裡に開催できなかった．特にデータの収集と提供に協力していただいた方々に感謝したい．データマイニングコンテストに多大な影響を与えたノースキャロライナ大学シャーロット校のJan M. Zytkow教授に感謝し，併せて同氏の急逝を悼みたい．

参考文献

1) Beilken, C. and Spenke, M.: Visual, Interactive Data Mining with InfoZoom - the Medical Data Set, Workshop Notes on Discovery Challenge, University of Economics, Prague, pp.49-54 (1999).
2) Brodley, C. E. and Smyth, P.: Applying Classification Algorithms in Practice, Statistics and Computing, 7 (1), pp.45-56 (1997).
3) Fayyad, U. M., Piatetsky-Shapiro, G. and Smyth, P.: From Data Mining to Knowledge Discovery: An Overview, Fayyad, U. M. et al. (eds.), Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, Menlo Park, Calif., pp.1-34 (1996).
4) McLachlan, G. J.: Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York (1992).
5) Mitchell, T. M.: Machine Learning and Data Mining, Comm. ACM, 42 (11), pp.31-36 (1999). [邦訳：機械学習とデータマイニング, CACM日本語版, 1 (1), pp.7-12 (2000).]
6) Suzuki, E. (ed.): Proc. Int'l Workshop of KDD Challenge on Real-World Data, (http://www.slab.dnj.ynu.ac.jp/challenge2000) (2000).
7) 山口高平編: 特集「共通データによる知識発見手法の比較と評価」, 人工知能学会誌, 15 (5), pp.750-797 (2000).
8) Zytkow, J. M. (ed.): Machine Discovery, Kluwer, Dordrecht (1997).
9) (第5回共通データからの知識発見), 人工知能学会研究会資料 SIG-KBS-A002, 人工知能学会, pp.45-108 (2000).

(平成13年3月31日受付)