特　集

機械学習とデータマイニング

Machine Learning and Data Mining

元田　浩*　鷲尾　隆*
Hiroshi Motoda　Takashi Washio

*　大阪大学産業科学研究所
Institute of Scientific and Industrial Research, Oaska University, 8-1 Mihogaoka, Ibaraki, Osaka 567

19YY 年 MM 月 DD 日　受理

Keywords: Machine Learning, Data Mining, Knowledge Discovery, Induction, Feature Selection, Scientific Discovery

Summary

Decision tree learning and rule induction are the main components of the current data mining (DM) algorithms, but machine learning (ML) is not a part of DM, it has its own identity just as statistics does. In this article, some of the exisiting techniques of ML that are useful for DM are reviewed and discussed. These include attribute selection methods, three learning paradigms (decision tree learning, inductive logic programming, graph-based induction) and scientific discovery. There is no doubt that ML forms the core of DM, but the current algorithms, although being successfully applied to real world problems, still suffer from computational complexity, especially for first order language learners that are most expressive.

1. はじめに

データマイニング (DM) は文字通り解釈すれば「データの中に埋め込まれている知識の発掘」を目的とするものである*1．機械学習 (ML) や統計処理の目的も知識の獲得である．これら 3 つの分野がデータベースからの知識獲得 (KDD) との関連でどのように位置付けられているかの通説を述べた後，DM における知識獲得を定式化し，それに ML の技術がどう関与するかを説明する．最初に属性の選択に関する幾つかの技術を紹介し，次に表現能力の異なる 3 種類の学習パラダイム: 1) 属性と値のペアで表現されたデータからの知識獲得の代表例である決定木学習，2) 表現能力の強力な一階述語論理を基にした帰納論理プログラムによる関係の学習，3) その中間に位置する命題論理で表現されたグラフパタンの学習，の概要を紹介し，現状と見通しを述べる．最後にデータから公理論的な知識を発見する点で DM の範疇にある科学的知識発見に対する新しい手法を説明し，データの値の他にデータの測定行為・認知の意味論の重要性を指摘する．

2. データマイニング，機械学習，統計

KDD のプロセスは複数のステップからなり，少なくとも次の 5 つ*2が含まれる [Mannila 96]．1) 対象領域の理解，2) データの準備，3) パタン（知識）の発見，4) パタンの事後処理（視覚化，解釈など），5) 結果の活用．さらに，これらが繰り返されることが特徴的である．DM はこのうちパタンの発見に相当し，ML や統計はそこで使われている有用な技術，手法として位置付けられている．

DM にとって ML は単なる手法であるが，学習はもっと広い概念である．人間の学習は無意識下でも起こるし，スキルの学習のように言葉で明示的に表現出来ないようなものまで対象とする．発見と学習を対比して

*1　[Fayyad 96] によれば「KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in the data」と定義されている．

*2　[Fayyad 96] では 9 個のプロセスに細分されている．

見ればその違いがより明確になる．MLでは，通常，発見は観測データからの学習のうち高度のものと位置付けられている [Langley 86]．従って，MLからみればDMはその一部ということになる．とはいえ，これら3つはいずれも（実験）データから興味ある規則性，パターン，概念を見つけようとする共通点を有している．しかし，[Mannila 96] によれば3者は以下の点で微妙に違う．

DMでは知識発見のプロセスが強調されるが，MLでは帰納の部分のみが強調されることが多い．またMLではデータに潜む概念やメカニズムの存在が仮定されている．つまり，データはノイズや誤差で崩れているかもしれないが，その裏には学習すべき貴重なものが隠されているという期待が強い．一方，DMではデータが主役で，背後に意味のある構造の存在を必ずしも仮定してはいない．例えば，小売データではデータが全てであり，データから有益な情報が得られれればいいのであって，一般にはデータを完全に理解することに興味はない．獲得される知識に対する思い入れにも差がある．MLでは専門家でも難しい知識を獲得しようとするが，DMでは時間さえ十分あれば専門家なら獲得できる程度の知識を求めようとしている．扱うデータの量も両者を区別する要素となろう．DMではテラバイトのデータを扱うと宣伝されている．非常に大量のデータに対しては従来のMLの手法はそのままでは使えない．統計ではデータマイニングという言葉は永らく，それも明確な仮説なしに闇雲に探索するという悪い意味で，使われて来た．これに対して探求的データ解析 (exploaratory data analysis) という言葉があり，DMに近い．データ主導の探索と言う意味合いが含まれる．最近の統計の分野ではモデル推定（モデルを仮定してそのパラメータをデータから最適化）からモデル選択（構造まで探索の範囲に取り込む）に流れが移りつつあり，DMとの相性がよくなって来た．データの不確かさの扱いに関してはMLもDMも統計に負っている．DMではMLの手法がよく使われるが，統計にはMLの手法はあまり使われない．統計でも量の問題は解決されていない．数10万のパラメータの推定はまだ困難である．

3. データマイニングから見た知識獲得

Imielinski[Imielinski 95] は今日のDMの研究状況は1960年代のデータベース (DB) の研究状況と同じであると述べている．当時はすべてのデータ管理のアプリケーションプログラムはアドホックであった．その後の関係モデルと強力な高級問合せ言語の登場で，初めて一般ユーザはデータベースの構造を意識しないで自分の望むものを簡単に問い合わせることが出来るようになった．この間30年かかっている．DMにおける知識獲得はこの問い合わせの中身が高級になったものであり，DBと同じ歴史を歩むことが予想される．

問合わせ言語の観点からDMを一般的に定式化することは容易である．

$$DM(D,P) = \{p \in P \mid p \text{はデータ} D\text{中に頻発する興味深いパターン}\}$$

相関ルールの場合はPは$A \Rightarrow B$の形のルールになり，確信度が十分大きいことが興味深さの指標となろう．決定木の場合にはPは決定木，予測誤差が所望の値以内であることが興味深さの指標となろう．上のDM述語を実現するアルゴリズムは図1と書ける．

```
C : Pの要素から生成した初期パターン;
While C ≠ φ do
    For each p ∈ C
        Dの中の p の頻度を評価
        If p は頻発 then F := F ∪ {p}
        end_if
        C := Fの要素から新たに生成したパターンの候補
    end_for
end_while
Fを出力
```

図1 データマイニングのアルゴリズム

発見のための問合わせ言語の仕様，コンパイル，実行方式の検討が今後の大きな課題である．このような方向の第一歩としてルール問合わせ言語が試作されている [Imielinski 96]．

4. 機械学習が貢献できる技術

4・1 属 性 選 択

特殊な場合を除き，一般のデータベースには，とくに既存の複数のデータベースからDMする場合には，当面の問題（例えば分類規則の学習）に関係のない属性や冗長なものも多数含まれているし，さらにノイズも混入している．必要な属性を選択する問題はデータクリーニングの重要な一部であり，MLでは多くの研究がある．属性の数をNとすると属性パターン（部分集合）の数は2^Nとなる．Nが大きくなると組合わせは天文学的な数になり，不要な属性をいかに効率良くふるい落とせるかが死命を制する．属性選択には大きく分けてフィルター法とラッパー法の2つがある．前者は属性選択の評価に適当な代替基準を用い，後者は学習評価結果そのもの（例えば分類誤差）を用いる．当然，後者の方が選択の精度はいいが，学習アルゴリズムを内蔵

するので処理時間の点から実用的ではない．ここではフィルター法について代表的なものを 2，3 紹介する．代替評価基準としては，a_1) 距離尺度や，a_2) 整合性尺度がよく用いられる．探索法に関しては，b_1) ヒューリスティックスを採用するもの，b_2) 完全性を保証するもの，b_3) ランダムに探索するものに分類される．以下の例は全て属性からクラスを同定する分類問題を対象にしている．(a_1, b_1) に属するものに $Relief$[Kira 92] がある．ある事例とそのニヤミス（クラスが違う，属性パターン間の距離が最小な事例）を区別する属性の方が，その逆の，その事例とニヤヒット（クラスが同じ，属性パターン間の距離が最小な事例）を区別する属性より重要とのヒューリスティックスを用いている．アルゴリズムを図 2に示す．$Relief$はノイズに強く，混在属性（連続数値，離散数値，名義）にも適用可能であるが，冗長性に弱く，クラスはバイナリーに限定されている．(a_1, b_2) に属するものでは分岐限定法 ($B\&B$) がよく

```
Relief(S);
  全ての重みを 0 に初期化
 For j = 1 to No_of_Sample
    ランダムにデータを 1 つ選択;
    ニヤヒット (hit) とニヤミス (miss) を検索;
    For 全ての属性 f_i
       W_i := W_i − δ(x_ji, hit_ji)^2 + δ(x_ji, miss_ji)^2
    end_for
 end_for
 W_i := W_i/No_of_Sample
 For 全ての属性 f_i
    If W_i > 閾値 then S_0 := S_0 ∪ {f_i}
    end_if
 S_0 を出力
```

図 2　Relief のアルゴリズム

知られている．評価基準が単調な場合（属性パターンの部分集合は親よりよくなることはないという性質）には全解探索をしないでも完全性が保証される．この方法では，望ましい属性数 Mを指定し，評価が最大の属性集合を求めるか，評価値の下限δを指定して，それを満足する最小の属性集合を求めることになる，前者のアルゴリズムを図 3に示す．全属性集合から 1 つづつ属性を削除する後方探索が基本である．多くの評価基準は単調性を満足しないので，この要求を緩和した改良版がある．実績では連続数値，離散数値は扱えるが名義属性がうまく扱えない．ノイズに対する評価は不明である．

(a_2, b_2) に属する代表的なものに $Focus$ がある．空の属性集合から属性を 1 つづつ追加する前方探索で，整合性が保持できる（全ての属性値が同じならクラスも同じであること）最小の属性集合を求めるものである．アルゴリズムを図 4に示す．非常に簡単な方法であるが，連続数値やノイズが扱えない．　(a_2, b_3) に属

```
B&B(S);
  For S中の各属性 f_i
      S_i := S − f_i
  end_for
  For 全ての S_i （評価値 U(S_i) の大きい順に）
      If U(S_i) > U_0
         If |S_i| = M then S_0 := S_i, U_0 := U(S_i)
         else S_0 := B&B(S_i)
         end_if
      end_if
  end_for
  S_0 を出力
```

図 3　分岐限定法のアルゴリズム

```
Focus(S);
  For i = 1 to No_of_Attribute
     For サイズ i の S の各部分集合 S_i
        If 不整合度= 0
           解候補部分集合:= S_i;
           Goto out
        end_if
     end_for
  end_for
  label : out
  解候補部分集合 S_i を出力
```

図 4　Focus のアルゴリズム

するものに LVF[Liu 96] があるが，誌面の都合で割愛する．(a_2, b_1), (a_1, b_3), (a_2, b_3) の組合わせは著者等の知る限りない．

筆者等は最近 $B\&B$に評価尺度として単調性を有する不整合度を導入して全属性集合に対する不整合度δと同じ不整合度を有する最少属性集合を求める ABBを考案した．$UC\ Irvine$ のデータを用いた性能評価結果を表 1に示す．不要属性の削除により探索空間が大幅に減少している．分類精度は $ID3$ を用いて 10-fold cross validation で評価した値である．属性選択によって誤差の増加は見られない．この他にも素性の分かったデータで確かに不要属性が削除されることが確認されている．$Focus$ と同じく連続数値が扱えない．

表 1　ABBによる属性選択の効果

データ	探索範囲		属性数		分類誤差 (%)	
	前	後	前	後	前	後
WBC	2^9	188	9	4	5.8	5.4
LED-7	2^7	9	7	5	0.0	0.0
Letter	2^{16}	1971	16	9	28.1	27.6
LYM	2^{18}	82,156	18	6	23.9	22.4
Vote	2^{16}	301	16	8	2.3	2.1

4・2　帰納学習の手法

〔1〕 決定木の学習

現在 DM で相関ルールについでよく用いられている，分類のための学習手法である．ML では歴史も古く多く

の手法が提案されている．分類のアルゴリズムとしては基本的には図5に示す分岐征服(Divide-and-Conquer)アルゴリズムと図6に示すカバーリングアルゴリズムの2種類がある．ここでは決定木を学習する代表例として前者のアルゴリズムを採用している$ID3$[Quinlan 86], $C4.5$[Quinlan 93] の要点を紹介する．いずれも逐次学習機能はない．

```
DT(D)
  属性 A_j の選択
  データ D を属性 A_j で部分集合 D_i に分類
  For 各部分集合 D_i
    While クラスの数 > 1
      DT(D_i)
    end_while
  end_for
```

図5　分岐征服アルゴリズム

```
CV(D)
  正例の幾つかをカバーする AND条件を発見
  そのクラスの他の正例をカバーするよう一般化
  負例がカバーされれば，それを排除するよう特殊化
  上の条件を規則の OR条件に追加
  カバーされるデータ D_c を削除: D_r = D − D_c
  While D_r ≠ φ
    CV(D_r)
  end_while
```

図6　カバーリングのアルゴリズム

目標はデータに内蔵する本質をつかんだ決定木の学習であるが，有限のデータからそれを実現することは困難である．データに矛盾がない場合，正しい決定木は多数あるが，そのうち簡潔な決定木ほどいいことが多くの実験でも確認されている．このような性質を持った決定木を学習するために，頻度情報を利用して属性の重要度を推定する．簡単のためクラスは2値（正例Pと負例N）とし，その個数をp，nとする．クラスを同定するのに必要な平均情報量は

$$I(p,n) = -\frac{p}{(p+n)}\log_2\frac{p}{(p+n)} - \frac{n}{(p+n)}\log_2\frac{n}{(p+n)}.$$

データDを属性fの値$\{f_1, f_2, ..., f_v\}$で分類して出来る部分集合を$\{D_1, D_2, ..., D_v\}$とすると，分類後の平均情報量は

$$E(f) = -\sum_{i=1}^{v}\frac{p_i+n_i}{(p+n)}I(p_i,n_i).$$

$ID3$ では分割することによる平均情報量の利得（gain）が最大になる属性fを順次選択する．

$$gain(f) = I(p,n) - E(f)$$

この基準では値の種類vが大きい属性が選ばれやすいので，$C4.5$ではこれを分割そのものに必要な情報量DIで規格化した利得比(gain ratio)を用いている．

$$DI(f) = -\sum_{i=1}^{v}\frac{p_i+n_i}{p+n}\log_2\frac{p_i+n_i}{p+n}$$

属性値が連続数値の場合は利得比が最大となる値で分割する．$C4.5$は2分割しかしないが，これを再帰的に繰り返すことは可能である．

データに矛盾がある場合は同じ属性パターンについてPである事例とNである事例が混在するが，数の多い方のクラスを採用する（離散真理値の場合の最小期待誤差を与える）．属性の値が未知の場合は，1) 値が未知の事例の属性を値が既知の他の事例の頻度分布から推定，2) 決定木を逆に利用して値を推定，3) unknownという新たな属性値を作成する方法などがあるが，いずれもあまり良い結果が得られない．$C4.5$では値を求めないで利得比を計算する．値が未知のデータは$gain$に寄与しないので，値が既知のデータから計算した$gain$を既知データの割合で割引き，DIは値が未知という属性を1つ追加して計算する．属性値が未知のデータは既知データの頻度分布を重みとして各分岐に分類される（既知のデータは重み1で分類）．従って，パスが複数になり，各パスで計算されるクラス毎の重みの和が最大のものが推定値となる．

さらに$C4.5$では分類誤差の期待値が最小になるよう決定木をプルーニング（枝刈）している．また決定木からルールへの変換も機械的にできる（説明は省略）．

この方法は計算量も少なく，かなり大量のデータに適用可能である．また，ノイズや欠損データに対する対策もよく研究されている．ただ，属性表現では表現力が弱く，既知の背景知識をうまく反映できないことが欠点である．

〔2〕 関係の学習

属性表現では，複数の関係にまたがるパターンを簡潔に表現できない．このような関係を表現するにはより強力な言語が必要である．現在知られている手法では一階述語論理を用いた帰納論理プログラミング(ILP)が最強であろう．この方法は探索空間が非常に広いので，表現能力を制限（MLではバイアスと呼んでいる）したり，探索制御にヒューリスティックスを用いるなどの対策が必要となる．しかし，領域知識や学習した知識を背景知識として利用できる大きなメリットがあり，MLの分野でも急速に注目を浴びている．

通常のILPの問題設定は，学習したい目標概念（述語p）の正例D^+と負例D^-，ならびに背景知識B（pの定義に使ってよい他の述語q_iや事実）から，与えられ

た言語 L を使って，p の定義 H（$\forall d \in D^+ : B \wedge H \vdash d, \forall d \in D^- : B \wedge H \not\vdash d$) を学習するものである．$H$ の帰納には種々の方法が提案されているが，基本的には以下の 5 つの技術が使われる [Džeroski 96][*3]．

相対最小汎化 (Relative least general generalization, rlgg): 2 つの節が正しいなら，それらの一般化の内で最も特殊なもの (lgg) も正しいであろうとするもので，背景知識として成立する事実 Kの下での p の事例 p_1, p_2の rlgg を求めることを繰り返すボトムアップ的な手法．$GOLEM$で採用されている．結果が特殊すぎる傾向がある．

$$rlgg(p_1, p_2) = lgg((p_1 \leftarrow K), (p_2 \leftarrow K))$$

逆導出 (Inverse resolution): 導出原理 ($A, \neg A \vee B \vdash B$) を逆に使い ($A, B \rightarrow \neg A \vee B$)，定数を変数に置き換える逆代入により汎化するもの．この操作は absorption (V) operator と呼ばれているが，これを 2 つ組合わせた W operator と呼ばれているものなど幾つかの逆導出の操作がある．$CIGOL$ や $PROGOL$ で採用されている．

特殊化のためのラティス探索: 一般的な節から出発して負例がカバーされなくなるまで特殊化（変数を項に置換，ボディーに述語の追加など）する仕組み (カバーリングアルゴリズムの一種）．仮説空間はθ包摂 ($C\theta \subseteq D$のとき Cは Dより一般的）と呼ばれるラティスを構成する．$FOIL$ や $FOCL$ で採用されている．このトップダウン探索にはヒューリスティックス ($FOIL$ では情報量基準を用いている）が使われる．

ルールモデルの採用: 学習したい節の形を与える．例えば，$P(X,Y) \leftarrow R(X), Q(Y,X)$ に制限する．それでも Bの述語とルールモデルの述語のマッチ（2 階），引数のマッチ（1 階）の全てを試すので探索空間は膨大となる．$MOBAL$ で採用されている．

命題表現への変換: 変数のタイプの制限内で背景知識の述語の引数に目標述語の引数の全ての組合わせを代入し，それらを新しい属性と考え，正例，負例の引数から各属性の値を求め，決定木などの属性表現の扱える学習アルゴリズムを用いルールを学習する．もとまったルールを節に変換する．この方式は表現能力に強い制約がある（再帰が駄目，ボディーの変数は全てヘッドにも現れるなど）がノイズを扱うことが出来る利点がある．$LINUS$に採用されている．

最近の ILPの進展は目覚ましく，一度に複数の述語を学習 [Zhang 96] したり，数値的な制約条件を取り入れることが出来るものもある．KDD への応用も盛んで，蛋白質の 2 次構造の予測，河川の水質診断，化学物質の突然変異性の予測など，まだデータの規模は大きくはないが，問題をうまく設定することにより実世界のデータに適用した例も増加している．しかし，処理効率の点でまだまだ問題が多い．

〔3〕 グラフパタンの学習

関係の中にも述語論理ほど強力な表現力がなくても表わせるものは多い．多くの事例パターンは有向グラフで表現できる．これは命題論理で特殊な関係を表現したことに相当する．筆者等はグラフ構造の中から繰返し現れるパタンを発見する GBI法を提案し，分類規則の学習，概念の学習，抽象化や近似による概念の構造化，マクロ規則の学習などが可能であることを示した [Yoshida 95, Yoshida 94]．図 7に示すように逐次ペアをチャンクしながらパターンを拡張して行くことが特徴である [吉田 97]． グラフ中のサブグラフの探索は NP-hard な問題であるので，ノードから出るリンクの順番に意味を持たせ，順番も含めて同じものしか同じものとみなしていない．図 8にアルゴリズムを示す．ペアノードの選択基準は頻度最大のもの，他と比べて特定しやすいものなど「興味深いパターン」を取り出すのに有効なものを採用する．停止条件も選択基準に応じたものを採用する．

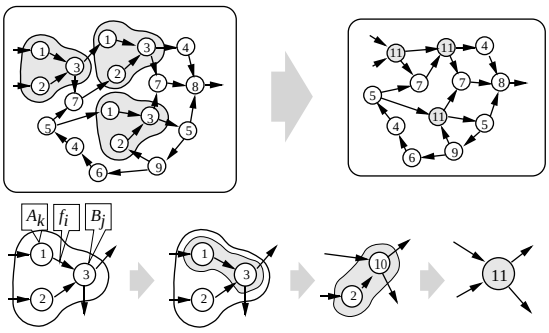

図 7 グラフに基づく帰納学習法 GBIの原理

$GBI(G)$
　ペアノード (A_k, f_i, B_j) の選択
　ペアノードをチャンク: c
　$C := C \cup \{c\}$
　G_c :=グラフ G の縮約
　$While$ 停止条件を未達
　　$C := C \cup GBI(G_c)$
　end_while
　Cを出力

図 8 GBI のアルゴリズム

分類規則の学習の場合にはルートノードをクラス，それにリンクするノードをグラフ構造をした属性とみ

[*3] この項で引用したプログラムに関する文献は [Džeroski 96] にある．誌面の都合で割愛する．

なし，ルートノードからリンクを辿りながらチャンクするような制御を入れる．但し，選択する対象は属性ではなくペアノードなので，属性とその値を選択する必要がある．この場合でも $ID3$ などと同じ情報量に基づく基準を採用することができる．

値が $A_k(k=1,K)$ のノードの数を n_k，$N=\sum_k n_k$ とすると選択前の情報量は

$$I(\overline{n})=\sum_k \frac{n_k}{N}\log_2\frac{n_k}{N},$$

属性 f_i の値 B_j で分類した後の情報量（$N_{ij}^{Yes}, N_{ij}^{No}$ は分類後の各部分集合の要素数）は

$$E(Attr_i,B_j)=\frac{N_{ij}^{Yes}}{N}I(\overline{n_{ij}^{Yes}})+\frac{N_{ij}^{No}}{N}I(\overline{n_{ij}^{No}}).$$

この差が最大となる属性 f_{i_0} の値 B_{j_0} で分類すればよい．

筆者等はこのアルゴリズムを計算機コマンド履歴をグラフ表現して，次のコマンドを予測する問題に適用した [Yoshida 96]．$ID3$ のような属性表現では時系列情報しか取り込めないが，グラフ構造を用いることによりファイルの入出力関係も依存情報として取り込める．その結果，予測精度が 35%から 75% に向上した．この方法はまだ連続数値が扱えない．

ここで紹介した 3 種類の学習法を比較すると知識表現能力と探索空間の広さでは「決定木 $< GBI < ILP$」であるが，計算の効率性は「決定木 $> GBI \gg ILP$」となる．とくに，GBI と ILP の間の差が大きいので，GBI は属性表現では解けないが，ILP ほどの表現力が要求されない問題には効果的である．

5. 公理論的な知識の発見

上で説明した学習法は，いずれも，データを説明する簡潔な表現を求めるものであるが，学習した結果が単なる実験式なのか，それとも第一原理であるかの区別が出来ない．公理論的な知識を発見したい場合には，それを保証する枠組みが必要となる．本章では「科学的知識発見」と呼ばれている，物理法則などの第一原理法則式を自動発見する新しい手法を紹介する．実験データから法則を発見するシステムの代表的なものに $BACON$ がある [Langlay 87]．同様な原理により更に拡張された機能を持つシステムもその後幾つか開発されている [Falkenhainer 86]．これらは基本的に，実験的測定において 2 変量毎の間に比例関係や加算関係，その他の種々の関係を仮定し，探索し，それらを繰り返し全体の関係を求めるものである．この方法は，実験データが表す多様な多変数関係の候補を導出できる利点があるが，正しさの保証に関する上述した問題を抱えている．

公理論的原理に基づき第一原理法則式を自動発見する研究は大きく 3 つに分類される．1 つ目は実験対象が有する対称性に関する知識を用いるものである [Ishida 95]．例えば，振り子は鏡に映しても挙動は不変であり，このような知見に基づき支配法則式の形式を絞り込んでいく．2 つ目は，次元解析手法 [Bridgman 22],[Buckingham 14] を基に，変量関係を探索するものである [Falkenhainer 86]，[Kokar 86]．例えば，ある物体の持つ運動エネルギー $E(J\equiv kgm^2/s^2)$ と質量 $m(kg)$，速度 $v(m/s)$ の間には，3 者の単位次元が整合するためには $E=cmv^2$(但し c は定数) という関係しかあり得ない．しかし，物理学の領域以外では，必ずしも対称性や数量の単位次元が明からではないことが多く，$BACON$ の方法にくらべ適用領域が限られる．3 つ目のものは，単位次元の公理よりも更に普遍的に成立する数量のスケールタイプの公理を用いる方法である [Washio 97]．この方法はより広い問題領域へ適用可能であり，以下にその概要を紹介する．

数量のスケールタイプの公理に関する研究は，公理論的測定論という研究分野で大きく進展した．Stevens は測定過程を「一定の規則によって対象や事象に数を割り当てること」と定義し，かつ異なる規則の下で数が割り当てられれば，異なる種類の数量のスケールと測定が導かれると主張し，比例尺度や間隔尺度，順序尺度，名義尺度などを定義した [Stevens 46]．特に前 2 者は定量的数量の主要なものである．比例尺度は質量や圧力，周波数，金額など，絶対原点基準を有し 2 つの測定量の比率が不変な量である．間隔尺度は摂氏や華氏で測った温度やエネルギー，時刻，音程など，人間が任意に設定した基準原点から測った量であり 2 つの測定量の差のみが実質上の意味を持つ量である．また別に，絶対的な原点と目盛り間隔を有する絶対尺度が存在する．これは 2 つの棒の長さの比のように一般に無次元量と呼ばれる量である。

このような定式化を受けて，Luce は各種スケールの 2 つの数量が絶対尺度を介さないで直接の依存関係を持つなら，それらには両者のスケールの種類に依存する基礎的関数のみしか許容されないことを非負の数量について証明した [Luce 59]．筆者等はこれを負値領域にまで拡張し，表 2の結果を得た [Washio 97]．

一方，上記の公理論的測定論とは別に，比例尺度の数量のみからなる第一原理式については，次元解析の *Buckingham* Π-*theorem* と呼ばれる重要な定理が成立する [Buckingham 14]．これによれば，$\phi(x,y,....)=0$ が前述の単位次元の制約の意味で第一原理を表す 1

表 2　尺度の性質を満たす可能な関係式

尺度の種類			
No.	独立変数	従属変数	可能な関係
1	ratio	ratio	$u(x)=\alpha_*\|x\|^\beta$
2.1	ratio	interval	$u(x)=\alpha\log\|x\|+\beta_*$
2.2			$u(x)=\alpha_*\|x\|^\beta+\delta$
3	interval	ratio	不可能
4	interval	interval	$u(x)=\alpha_*\|x\|+\beta$

1) 表記 α_*,β_* はそれぞれα_+,β_+ for $x\geq 0$ and α_-,β_- for $x<0$ を表す.

つの完全な式であれば，$F(\Pi_1,\Pi_2,...,\Pi_{n-r})=0$ という形式に書換え可能である（n は ϕ の引数の数，r は $x,y,z....$の単位に含まれる長さ $[L]$、質量 $[M]$、時間 $[T]$ のような基本単位次元の数）．またすべての i について，Π_iは無次元量であり，かつ $x,y,z....$の関数$\Pi_i=\Pi_i(x,y,....)$ である．Fを ensemble，Π_iを regime と呼ぶ．また，各regime Π_iは$product\ theorem$ によって，$f=Cx^ay^bz^c....$という形式を持つことも知られている（$C,a,b,c,...$ は定数）．筆者等は，前述の公理論的測定論の比例尺度と間隔尺度に関する議論を基に，これらの定理を以下のように拡張した[Washio 97]．

Theorem 1（拡張 Buckingham Π-theorem） $\phi(x,y,....)=0$ が 1 つの完全な式であるならば，それは以下のような形式に書き換え可能である（n はϕの引数の数，r,s は $x,y,z....$の基本単位次元の数及び間隔尺度の基本原点の数）．

$$F(\Pi_1,\Pi_2,...,\Pi_{n-r-s})=0$$

またすべての i について，Π_iは無次元量である．

Theorem 2（拡張 product-theorem） 1 つの $regime$ を構成する数量の集合 $Q=\{x_1,x_2,...,x_m\}$ があり，そのうちの幾つかは間隔尺度，他は比例尺度であるとする．この時，無次元量Πは以下の 2 式の何れかの形式を取る．

$$\Pi(x_1,x_2,...,x_m)=(\prod_{x_h\in R}|x_h|^{a_h})\times \prod_{I_i\subseteq I}(\sum_{x_j\in I_i}b_{*j}|x_j|+c_{*i})^{a_i} \qquad (i)$$

$$\Pi(x_1,x_2,...,x_m)=\log\{(\prod_{x_h\in R}|x_h|^{a_h})\times \prod_{I_i\subseteq I}(\sum_{x_j\in I_i}b_{*j}|x_j|+c_{*i})^{a_i}\} +\sum_{x_k\in I'\subseteq I}b_k|x_k|+c \qquad (ii)$$

ここで，Rは Q 内の比例尺度量の集合，Iはそれ以外，即ち間隔尺度量の集合である．また，全ての I_iについて $I'\cap I_i=\phi$である．

筆者等はこの 2 つの定理に基づき，実験的測定データと各数量のスケールタイプ（間隔尺度か比例尺度か）を与えると，第一原理として可能な法則式を定理が許容する関数関係式の範囲内で探査するシステムを開発した．以下にその適用例として，1 つの $regime$ である理想気体の状態方程式を導く場合を説明する．この $regime$ は，圧力 p，体積 v，質量 m という比例尺度量と温度 t という間隔尺度量 (絶対温度でない場合) から成る．ここで p，v，m の値が正である場合を考えると，定理 2により 2 つの候補が示される．

$$\Pi=p^{a_1}v^{a_2}m^{a_3}(bt+c)^{a_4},$$
$$\Rightarrow p^{a_1}v^{a_2}=\Pi m^{-a_3}(bt+c)^{-a_4}.$$
$$\Pi=\log p^{a_1}v^{a_2}m^{a_3}+bt+c,$$
$$\Rightarrow p^{a_1}v^{a_2}=m^{-a_3}\exp(-b_1t-c+\Pi).$$

前者は温度単位が絶対温度に限らない場合の理想気体状態方程式を正しく表している．一度 $regime$ の候補式が決定されると，正しい式とその係数の値は，測定データへのフィッティングによって求められる．この場合には，前者が正しい式と判明し各係数の値が定められた．各数量の単位次元やデータの情報なしに，非常に狭い範囲にまで候補式が絞り込まれることに注目されたい．これは $regime$ が 1 つのみの場合の例であるが，3 つのトランジスタからなる電子回路増幅器 (17 変量かつ 7$regime$ を含む) のような大規模なシステムに関しても同様に第一原理式を導くことに成功している．

6. おわりに

機械学習の諸技術をデータマイニングの観点から概説した．日々新しい進展があり，学会の度に新しい手法が提案されている．属性表現を用いた決定木やルールの学習はかなり成熟した段階にあるが，それでも細かな改良が続けられている．ILPの分野はまだ日が浅いが，表現能力の高さを武器に，多くの期待が寄せられており，ノイズ対策や数値制約の取り込など従来手法との融合が進展している．しかし，現段階ではテラバイト級のデータベースに対しては相関ルールの発見のように有効なアルゴリズムはない．学習アルゴリズムの改良は探索空間の枝刈が中心で，2 次記憶装置へのアクセス最小化やメモリー管理の最適化までは目が向けられていない．この意味では，非常に大量のデータに対する課題は山積みしていると言えよう．属性選択は今後ますます重要になるであろう．新しい属性を既知の属性から帰納的に構成することも重要で，この点でもILPは期待が持てる．統計処理の技術は単独でも ML の技術の一部としてでも，これからも基盤技術としてその地位を保つであろう．データベースの観点から言えば，現在は分離しているデータベース管理システムとデータマイニング (ML，統計処理含めて) のアプリケーションプログラムはいずれ融合されるであろう．しかし，それにはデータベースの発展を考える

とかなりの時間がかかることが予想される．

◇参 考 文 献◇

[Buckingham 14] Buckingham, F.: On Physically Similar Systems; Illustrations of the Use of Dimensional Equations. *Phys. rev.*, Vol. IV, pp. 345–376, 1914.

[Džeroski 96] Džeroski, S.: Inductive Logic Programming and Knowledge Discovery in databases. In U. M. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 117–152. AAAI Press, 1996.

[Falkenhainer 86] Falkenhainer, B. C. and Michalski, R. S.: Integrating Quantitative and Qualitative Discovery: The Abacus System. *Machine Learning*, Vol. 1, pp. 367–401, 1986.

[Fayyad 96] Fayyad, U. M., Piatesky-Shapiro, G. and Smyth, P.: From Data Mining to Knowledge Discovery: an Overview. In U. M. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pp. 1–34. AAAI Press, 1996.

[Imielinski 95] Imielinski, T.: A Database View on Data Mining. In *Invited talk at the KDD'95 conf.*, 1995.

[Imielinski 96] Imielinski, T., Virmani, A. and Abdulghani, A.: Datamine: Application Programming Interface and Query Language for Data Mining. In *Proc. of The Second Int. Conf. on Knowledge Discovery and Data Mining*, pp. 256–261, 1996.

[Ishida 95] Ishida, Y.: Symmetry-Based Reasoning About Equations of Physical Laws. In *Working Papers of Ninth Int. Workshop on Qualitative Reasoning*, pp. 84–93, 1995.

[Kira 92] Kira, K. and Rendell, L. A.: The Feature Selection Problem: Traditional Methods and a New Algorithm. In *Proc. of The Ninth National Conf. on AI*, pp. 129–134, 1992.

[Kokar 86] Kokar, M. M.: Determining Arguments of Invariant Functional Descriptions. *Machine Learning*, Vol. 1, pp. 403–422, 1986.

[Langlay 87] Langlay, P. W., Simon, H. A., Bradshow, G. and Zytkow, J. M.: *Scientific Discovery; Computational Explorations of the Creative Process.* MIT Press., 1987.

[Langley 86] Langley, P. and Michalski, R. S.: Machine Learning and Discovery. *Machine Learning*, Vol. 1, No. 4, pp. 363–366, 1986.

[Liu 96] Liu, H. and Setiono, R.: A Probabilistic Approach to Feature Selection - A Filter Solution. In *Proc. of The Thirteenth Int. Conf. on ML*, pp. 319–327, 1996.

[Luce 59] Luce, R. D.: On the Possible Psychological Laws. *The Psychological Review*, Vol. 66, pp. 81–95, 1959.

[Mannila 96] Mannila, H.: Data Mining: Machine Learning, Statistics, and Databases. In *Proc. of The Eight Int. Conf. on Scientific and Statistical Database Management*, pp. 1–34, 1996.

[Quinlan 86] Quinlan, J. R.: Induction of Decision Trees. *Machine Learning*, Vol. 1, pp. 81–106, 1986.

[Quinlan 93] Quinlan, J. R.: *C4.5: Programs for Machine Learning.* Morgan Kaufmann, 1993.

[Stevens 46] Stevens, S. S.: On the Theory of Scales of Measurement. *Science*, Vol. 103, pp. 677–680, 1946.

[Bridgman 22] Bridgman, P. W.: *Dimensional Analysis.* Yale University Press., 1922.

[Washio 97] Washio, T. and Motoda, H.: Discovery of First Principle Based on Data-Driven Reasoning. In *Proc. of the First Pacific-Asia Conf. on Knowledge Discovery and Data Mining*, pp. 169–182, 1997.

[Yoshida 94] Yoshida, K., Motoda, H. and Indurkhya, N.: Graph-based Induction as a Unified Learning Framework. *J. of Applied Intelligence*, Vol. 4, pp. 297–328, 1994.

[Yoshida 95] Yoshida, K. and Motoda, H.: Clip: Concept Learning from Inference Pattern. *J. of Artificial Intelligence*, Vol. 75, No. 1, pp. 63–92, 1995.

[Yoshida 96] Yoshida, K. and Motoda, H.: Automated User Modeling for Intelligent Interface. *Int. J. of Human Computer Interaction*, Vol. 8, No. 3, pp. 237–258, 1996.

[Zhang 96] Zhang, X. and Numao, M.: Efficient Multiple Predicate Learner Based on Fast Failure Mechanism. In *Proc. of The Fourth Pacific Rim Int. Conf. on AI*, pp. 35–46, 1996.

[吉田 97] 吉田健一, 元田 浩: 逐次ペアに基づく帰納推論. 人工知能学会誌, Vol. 12, No. 1, pp. 58–67, 1997.

著 者 紹 介

元田　浩(正会員)

1965 年東京大学工学部原子力工学科卒業．1967 年同大学院原子力工学専攻修士課程終了．同年，日立製作所に入社．同社中央研究所，原子力研究所，エネルギー研究所，基礎研究所を経て 1995 年退社．現在，大阪大学産業科学研究所教授 (知能システム科学研究部門、高次推論研究分野)．原子力システムの設計，運用，制御に関する研究，診断型エキスパート・システムの研究を経て，現在は人工知能の基礎研究，とくに機械学習，知識獲得，知識発見などの研究に従事．工学博士．日本ソフトウェア科学会理事，人工知能学会理事，同編集委員，Knowledge Acquisition (Academic Press) 編集委員，IEEE Expert 編集委員を歴任．Artificial Intelligence in Engineering (Elsevier Applied Science) 編集委員，International Journal of Human-Computer Studies (Academic Press) 編集委員，日本認知科学会編集委員会委員，同常任運営委員．1975 年日本原子力学会奨励賞，1977，1984 年日本原子力学会論文賞，1989，1992 年人工知能学会論文賞受賞．人工知能学会，情報処理学会，日本ソフトウエア科学会，日本認知科学会，AAAI，IEEE Computer Society，各会員．

〈motoda@sanken.osaka-u.ac.jp〉

鷲尾　隆(正会員)

1983 年東北大学工学部原子核工学科卒業．1988 年同大学院原子核工学専攻博士課程終了．工学博士．同年，マサチューセッツ工科大学原子炉研究所客員研究員．1990 年三菱総合研究所入社．1996 年退社．現在，大阪大学産業科学研究所助教授 (知能システム科学研究部門、高次推論研究分野)．1987〜1988 年日本学術振興会特別研究員．原子力システムの異常診断手法に関する研究，定性推論に関する研究を経て，現在は人工知能の基礎研究，とくに機械学習，知識獲得，知識発見などの研究に従事．人工知能学会基礎論研究会幹事，計測自動制御学会知能工学部会運営委員．1987 年計測自動制御学会学術奨励賞受賞．1995 年人工知能学会全国大会優秀論文賞受賞，1996 年日本原子力学会論文賞受賞，1996 年人工知能学会研究奨励賞受賞．人工知能学会，計測自動制御学会，日本ファジイ学会，日本原子力学会，AAAI，各会員．

〈washio@sanken.osaka-u.ac.jp〉